JUNE
1961
VOLUME 25
NUMBER 3

JOURNAL OF CONSULTING PSYCHOLOGY

AMERICAN PSYCHOLOGICAL ASSOCIATION

JOURNAL OF CONSULTING PSYCHOLOGY

EDWARD S. BORDIN, Editor
University of Michigan

BOYD MCCANDLESS, Associate Editor
University of Iowa

Advisory Editors: John W. Atkinson, *University of Michigan*. Frank X. Barron, *University of California, Berkeley*. John M. Butler, *University of Chicago*. Allen T. Dittmann, *National Institute of Mental Health*. Seymour Fisher, *Baylor University*. Harrison G. Gough, *University of California, Berkeley*. William L. Hays, *University of Michigan*. Robert B. Malmo, *McGill University*. Paul E. Meehl, *University of Minnesota*. Daniel R. Miller, *University of Michigan*. Alan K. Rosenwald, *University of Illinois*. Saul Rosenzweig, *Washington University*. Julian B. Rotter, *Ohio State University*. Nevitt Sanford, *University of California, Berkeley*. Seymour B. Sarason, *Yale University*. Robert R. Sears, *Stanford University*. Edwin S. Shneidman, *Veterans Administration, Los Angeles, California*. Read D. Tuddenham, *University of California, Berkeley*.

The *Journal of Consulting Psychology* is the clinical journal of the American Psychological Association. It is devoted primarily to the publication of original research relevant to psychological diagnosis, psychotherapy and counseling, personality, and the dynamics of behavior. Case studies, theoretical contributions, descriptions of clinical techniques, and discussions of the professional problems of clinical psychology also appear in its pages.

Manuscripts submitted for *regular* or *early* publication and manuscripts for Brief Reports should be sent to the Editor: Edward S. Bordin, Department of Psychology, University of Michigan; Ann Arbor, Michigan. All manuscripts should be submitted *in duplicate* to facilitate referral to the Advisory Editors.

Style of Manuscripts. Manuscripts must be prepared in the style described in the *Publication Manual of the American Psychological Association (1957 Revision)*, obtainable for $1.00 from the Washington office of the APA.

Early Publication. Authors may secure the early publication of an acceptable article in the first issue to go to press after its acceptance by arranging for the article to be printed as extra pages, increasing the number of pages that subscribers receive. Information about charges for early publication may be obtained from the Managing Editor.

American Documentation Institute. To save printing costs, materials supplementary to articles may be deposited with the ADI. To obtain copies, order by Document Number as cited in the reference footnote, from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. The indicated fee must be remitted in advance. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Reprints. Authors receive fifty reprints gratis. Additional quantities may be ordered when returning proofs.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

Communications concerning subscriptions, change of address, claims for the nonreceipt of a number, advertising, and other business matters, should be sent to the American Psychological Association, Inc., 1333 Sixteenth St. N.W., Washington 6, D.C. Address changes must reach the subscription office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

The *Journal of Consulting Psychology* is published bimonthly in February, April, June, August, October, and December, the yearly volume comprising approximately 576 pages. The subscription per year is $10.00, foreign $10.50, single number $2.00.

Published by the American Psychological Association, Inc.
at Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth St. N.W., Washington, D.C.

Second-class postage paid at Lancaster, Pa.

© 1961 by the American Psychological Association, Inc.

JOURNAL OF CONSULTING PSYCHOLOGY

June 1961 Vol. 25, No. 3

Contents

ARTICLES

The Role of the Internship in the Research Training of the Clinical Psychologist: *Jules D. Holzberg* - 185

Comparability of Intelligence Quotients of Mental Defectives on the Wechsler Adult Intelligence Scale and the 1960 Revision of the Stanford-Binet: *Gary M. Fisher, Beverly A. Kilman, and Anna M. Shotwell* - - - - - - - - - - - - - - - - - - 192

Perceptual Size Constancy in Chronic Schizophrenia: *H. W. Leibowitz and Vladimir Pishkin* - 196

The Relationships between Individually Defined and Group Defined Social Desirability and Performance on the Edwards Personal Preference Schedule: *Alfred B. Heilbrun, Jr. and Leonard D. Goodstein* - 200

Behavior Problems of Middle Childhood: *Donald R. Peterson* - - - - - - - - - - - 205

Attribution of Traits and Emotional Health as Factors Associated with the Prediction of Personality Characteristics of Others: *Marvin Spanner* - - - - - - - - - - - - 210

Hostility Expression among Delinquents of Minority and Majority Groups: *Don L. Swickard and Bernard Spilka* - 216

The Use of Opposite Sex Scales as a Measure of Psychosexual Deviancy: *B. G. Rosenberg, B. Sutton-Smith, and E. Morgan* - 221

Test Anxiety and Performance under Stress: *Zanwil Sperber* - - - - - - - - - - - 226

Wechsler's Deterioration Ratio in Clinical Practice: *T. G. Crookes* - - - - - - - - 234

Vocabulary Deficit in Brain Operated Schizophrenics: *Roy M. Hamlin and Elaine F. Kinder* - 239

An Evaluation of the Northwestern Infant Intelligence Test, Test B: *Bernard B. Braen* 245

Judgments of Adjustment from TAT Stories as a Function of Experimentally Altered Sets: *Bernard Lubin* - 249

Orthopedic Disability as a Factor in Human-Figure Perception: *Aurelia Levi* - - - 253

TAT Performance as a Function of Anxiety and Coping-Avoiding Behavior: *E. Jerry Phares* - 257

Relationships between Descriptive Content and Interaction Behavior in Interviews: *Jeanne S. Phillips, Ruth G. Matarazzo, Joseph D. Matarazzo, George Saslow, and Frederick H. Kanfer* - 260

The Concept of Normality: A Reply to Freides: *Maurice Korman* - - - - - - - - 267

The Effects of Two Verbal Techniques on the Expression of Feelings: *Gustav Levine* - 270

BRIEF REPORTS

Anxiety in Verbal Behavior: An Intercorrelational Study: *Merton S. Krause* - - - - 272

Verbal and Perceptual Components in WISC Performance and Their Relation to Social Class: *John B. Marks and James E. Klahn* - - - - - - - - - - - - - - - - - - - 273

Sensory Deprivation and Its Relation to Projection: *Malcolm H. Robertson and Robert C. Martin* - 274

Psychiatric Outpatient Personality Patterns: *Mary Helen Tatom* - - - - - - - - - 275

Identification in Terms of Personal Constructs: Reconciling a Paradox in Theory: *Robert E. Jones* - 276

Brief Reports

The *Journal of Consulting Psychology* will accept Brief Reports of research studies in clinical psychology for early publication without expense to the author. The procedure is intended to permit the publication of soundly designed studies of specialized interest or limited importance which cannot now be accepted because of lack of space. Several pages in each issue will be devoted to Brief Reports, published in the order of their receipt without respect to the dates of receipt of the regular articles. Most Brief Reports appear in the first or second issue to go to press following their final acceptance.

An author who wishes to submit a Brief Report:

1. Sends the Brief Report, limited to one printed page and prepared according to the specification given below.

2. Also sends to the Editor a full report of the research study, in sufficient detail to give a clear account of its background, procedure, results, and conclusions, which will be filed with the American Documentation Institute to insure indefinite availability.

3. Prepares at least 100 mimeographed copies of the full report, which the author will send without charge to all who request it as long as the supply lasts.

4. Agrees not to submit the full report to another journal of general circulation.

Specifications

Brief Report. The Brief Report should give a clear, condensed summary of the procedure of the study and as full an account of the results as space permits.

To insure that the Brief Report will be no longer than one printed page, its typescript, including all matter except the title and the author's lines, must not exceed 85 lines averaging 42 characters and spaces in length. Set the typewriter margins for short lines of 42 characters, which are 3.5 inches long in elite typing, and 4.2 inches long in pica.

The manuscript of the Brief Report must be *double spaced* throughout. Except for its short lines, it follows the standard style of the 1957 revision of the *APA Publication Manual.* Headings, tables, and references are avoided or, if essential, must be counted in the 85 lines. Each Brief Report must be accompanied by a footnote in the style below, which is typed on a separate sheet and *not* counted in the 85-line quota:

[1] An extended report of this study may be obtained without charge from John Doe (giving the author's full name and address) or for a fee from the American Documentation Institute. Order Document No. —— from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $—— for microfilm or $—— for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

Extended report. Because the extended report is intended for photoduplication, and is not copy to be sent to a printer, its style should differ in several ways from that of other manuscripts: (*a*) The extended report should be typed with single spacing for economy in duplication. (*b*) Tables and figures should be placed adjacent to the text which refers to them. A caption should be typed below each figure. (*c*) Footnotes should be typed at the bottom of the page on which reference is made to them. In other respects, the full report is prepared in the style specified by the *Publication Manual.*

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 185–191

THE ROLE OF THE INTERNSHIP IN THE RESEARCH TRAINING OF THE CLINICAL PSYCHOLOGIST

JULES D. HOLZBERG[1]
Connecticut State Hospital

Training in clinical psychology has consisted of two aspects: (*a*) a minimum of 3 years at a university where the student receives a generic education in basic psychology, training in theory and techniques relevant to clinical psychology, and practicum experience in research; (*b*) at least 1 year of practicum experience (internship) in a clinical center to provide substance for the more academic aspects of the student's clinical training. Within this organization of clinical training, the university has been viewed as basically responsible for the training of the clinician as a psychological researcher, while the internship is generally conceived of providing only minimal support for the research aspects of a clinical student's training.

However, one may accept this basic responsibility of the university while at the same time questioning the adequacy of a conception that denies to the internship a significant responsibility in the training of the clinician as a researcher. The university is unquestionably better equipped from the point of view of facilities and personnel to provide training in the definition and application of psychological concepts, in the use of the library as a research tool, in experimental and research design, in the utilization of statistical analyses, in methods of sampling populations, and the numerous other requisite skills of the competent researcher.

However, the internship center has a significant role to play in the research training of the clinical student in at least two significant respects: (*a*) to sensitize the student to the meaningful research problems in the clinical field and to demonstrate how clinical skills and techniques may be adapted to research goals, while providing a framework that fosters research on clinical problems without destroying their clinical significance; and (*b*) to contribute to the building of a self-concept that effectively integrates the role of "researcher" and "clinician." The first of these aspects is by no means insignificant, for the university may, because of its frequent insistence on the absolute "researchability" of a problem, force the clinical student into a research area where the problem is "neat," but the clinical significance quite meager. This is not meant as a criticism of university training, since the author recognizes that the function of the university is to provide generic training in the development of research skills, and there may be real merit in focusing this training on "researchable" problems. We must, however, be prepared for one possible consequence of this training, i.e., this criterion of researchability may become so rigidly fixed that the clinician may never undertake research on a clinical problem, because it is rare indeed that a clinical problem can hope to attain the level of researchability that meets the standard set for students in the typical university.

The second research training function of the internship is related to the author's observations that clinical interns frequently come to the internship with a strong sense of inferiority as a researcher. Whether this is a function of the preselection of clinical students (in that students entering into the clinical area may do so because of intellectual and personality reasons that are incompatible with a research role) or whether this is a

[1] The author is indebted to Solomon E. Feldman, who read the manuscript and helped make it a more readable document. However, the author assumes full responsibility for the ideas expressed.

function of the absorption of attitudes toward the clinician as a research worker that unfortunately permeate many of our universities, the result often seems to be that when the clinical student enters into the internship, he does so with a self-concept that is relatively weak with regard to the integration of research as a practicing role.

The internship frequently does not help to counter such feelings of inferiority. It may fail to provide a "culture" which rewards the development of a research role for the clinician by encouraging other role behaviors, such as that of diagnostician and therapist. The practicing clinician in the internship center may further contribute to this self-concept by failing to provide an adequate model of the clinician as a research worker. He is usually not involved in research and may actually give voice to antiresearch attitudes, so that we have the vicious "cultural" cycle of the clinician, having learned to reject the role of research for himself, now fostering the same attitudes in the students sent to him for training.

The Acquisition of a Research Role[2]

I have been using two concepts which should be more precisely defined. One is that of "role" of a researcher, i.e., those actions which are performed or not performed to confirm the occupancy of the research role. The second, "self-concept," may be defined as the values used by a person to describe and evaluate himself, and here I am specifically referring to self-descriptions of the clinician with regard to values pertinent to being a research worker. These two concepts are functionally quite interrelated since attitudes toward the self affect the kind of role that will and can be played, and conversely, the success with which one plays or does not play a particular role will in turn affect the self-concept.

Two methods of role learning may be delineated, i.e., intentional learning and incidental learning. Intentional learning is explicit learning and the intentional learning of a research role occurs primarily in the university. In incidental learning, the individual adopts "the prevailing pattern." It is the author's thesis that incidental learning, much of which takes place in the clinical center, plays a significant part in the ultimate development of the self-concept of the clinician. Incidental learning occurs essentially through a process of identification (unconscious) or imitation (conscious). This places great emphasis on those in the emerging clinical psychologist's environment who have securely integrated the role of researcher into their professional self-concept. It is here that an important defect exists in the opportunities for learning a research role since the figures who securely integrate the research role into their self-concepts are frequently unavailable or minimally available in the internship. For the most part, there are figures who are diagnosticians, therapists, administrators, etc. The intern can *identify* with figures that are not in his immediate visual field, such as significant researchers in the field of psychology. However, we have perhaps failed to emphasize the fact that *imitation*, which may be more important in the development of the professional self-concept, does require the figure to be in the visual field.

This has been implicitly recognized in many clinical centers with the result that more clinical centers are today resorting to the use of research consultants who serve the important role of providing figures with a secure self-concept that incorporates the role of researcher. These are, for the most part, research psychologists from universities who visit the internship center at varying intervals, spending several hours on each visit and advising on theoretical conceptualizations, research designing, statistical analyses, and other issues pertinent to research. But even the use of consultants may be insufficient. Frequently, these are fleeting figures who appear irregularly, who may not actually become involved in research within the clinical center, and who are often not clinicians but are the prototypes of those academic figures who have originally contributed to feelings of inferiority about the clinician's research role.

[2] A number of the ideas presented in this paper were stimulated by Sarbin's (1954) excellent discussion of role theory.

This may be a considerable oversimplification. One may still have figures available who provide appropriate models for identification or imitation, but there may still not be a basic motivation for learning a research role. From where does the motivation to learn and practice the role of researcher stem? Such motivation may derive from factors such as financial rewards, status, advancement, and other extrinsic stimulants. However, we cannot ignore the fact that there may be more intrinsic factors operating to provide motivation for the learning of a research role. Redlich and Brody (1955) have emphasized research motivation as being derived from "powerful unconscious drives to gratify infantile curiosity and the wish for omniscience" (p. 234). Kubie (1953) has carried this further and has attempted to relate unconscious strivings to the various aspects of research such as the problem selected for study, the specific research techniques utilized, the hypothesis selected for verification, and other aspects of the research process.

In addition to learning the skills of research, including clinical research skills, and in addition to learning the role of researcher and effectively integrating it into one's self-concept, there is another type of learning that is a joint responsibility of the university and the clinical center. This is the learning of the ethics of research. I refer here to the obligations of the researcher to his subjects, to his profession, to the organization which employs him, to his collaborators, etc. Too often, one observes the researcher, be he clinical or not, whose competence in terms of knowledge and skills is obviously of a high order, but who has failed to integrate a valid system of ethical values into his research activity.

Problems of Role Enactment

One of the complex adjustment problems faced by the young clinician is the need to learn a multiplicity of roles, e.g., diagnostician, therapist, teacher, researchist, etc. At times, there may be conflict between certain of these roles, which may then result in the relinquishing of one of the roles in order to resolve the conflict. Where one of the roles is that of research, this frequently leads to the relinquishing of the research role because the clinical center frequently supports this role less than others.

Role conflict occurs if the individual occupies two positions simultaneously, and when the role expectations of one position are incompatible with the role expectations of the other. It is rare indeed that situations arise which present this type of role conflict as it pertains to the practice of research. Where one finds the verbalization of such role conflict, one must suspect that there exists another explanation, other than a real role conflict, since it is rare that the individual is expected to perform a research role and any other role simultaneously. The role of clinician and the role of research worker, even where these are successfully integrated, are not performed simultaneously, but are performed at different times, in different settings, in relation to different individuals, etc. The author views the multiple roles of the clinical psychologist not as mutually incompatible, but mutually complementary. The ability to successfully perform a multiplicity of professional roles permits one to enrich the particular role that the psychologist may be enacting at a particular time.

It has frequently been stressed that the attitudes and skills required for research are significantly different from those required for clinical activity. The problem here is not that there is a basic incompatability between research and clinical activity, but that certain types of research, usually of the experimental laboratory type, are strikingly different from what is involved in clinical practice. However, there are many problems which cannot at the moment be studied effectively in a strictly experimental laboratory manner, and here the clinical method, with its latitude, with its less standardized approach, with its greater subjectivity, can be marshaled to fulfill a research objective. The problem is one of adapting the clinical method so that it can become a useful research method. In fact, the author has stressed elsewhere the fundamental resemblance of the clinical and the scientific methods (Holzberg, 1957).

It is not being suggested that the clinical psychologist must enact all of his roles with the same degree of intensity or of involvement. It is the rare individual who can maintain intense affective involvement and intense expenditure of physical effort in all of the roles that he may be expected to perform. The clinician need not have the highest involvement in research, but somewhere in the repertoire of roles there must be integrated into the clinician the availability of the research role. Clinical psychologists vary in their enactment of the research role such that each may be located on a continuum extending from almost automatically enacting the role, to the other extreme where the clinician is very self-conscious of his own role enactment. The problem is how to make the clinician, and particularly the clinical student, become less self-conscious of the enactment of the research role.

To perform a role, the individual must clearly know the expectations of those with whom he interacts in this role. Unless the internship staff conceptualizes the intern's role as incorporating research, the intern's enactment of the research role may present many problems. Thus, he may act out a role which others may consider inappropriate, if not actually bizarre. Of crucial importance here are the overt acts of others, rather than verbalizations. If others are doing research, if research is being actively supported in the agency, this is a far more significant factor in revealing the attitudes toward a research role than mere verbalized support.

Where the professional structure demands a multiplicity of role performances as in clinical psychology, there must be "flexibility" in the individual, so that he may readily be able to shift from one role to another. It is apparent that this is one dimension of the self that may make for difficulty in the enactment of the research role, simply because the individual may be lacking in this capacity for flexibility. At the same time, the individual must possess sufficient capacity to erect barriers around a particular role being enacted, so that the role does not shift inappropriately.

Defensive Reactions to Research Role Conflicts

If the actions and qualities of a given role are congruent with the self-concept maintained by the individual, this enhances the probability that the individual will be able to perform in a way consistent with the role expectations. It is the thesis of the writer that the incongruence of the role of the researchist with the self-concept of the clinician is a central problem. This is in essence a situation of conflict, and like conflicts which are not directly resolvable, leads to tension and resort to various defensive operations. One of these is rationalization, which is best exemplified by the resort to the argument that the research role is incompatible with the clinician's role.

Another defensive reaction is that of projection. This manifests itself in destructive and nihilistic attitudes toward the research of others. This may take the form of the psychologist criticizing another reseacher's technical competence, his knowledge of the area, etc., all of which are projections of his own inadequacies.

Still another defense is that of denial—failing to recognize research problems or denying that problems others are posing are real research problems. This person seems to be saying: "If I do not recognize research problems, I cannot be expected to do research."

Some clinicians may resort to identification as a way of resolving the conflict. They may identify with strong clinical figures, such as psychoanalytic psychotherapists, who have considerable prestige in terms of their clinical endeavor. The clinician who can successfully make this identification may comfortably remove himself from a research role, since he can model his behavior after one who has acceptable status and prestige.

Some clinicians remain on the fringe of research. They serve as "advisers" to others, but never get involved in research. Frequently, they perform a significant role as advisers, which is testimony to the fact that they have certain competencies regarding research which should lead to research activity. These individuals are so blocked by a

sense of inferiority and the potential feeling that their research activity will not succeed that one finds this tragic discrepancy between genuine ability in research and an inability to function in a research role. At the other extreme is the "dilettante" who frequently does have some research interest and talent, but who wants to research every problem here and now. This frequently masks the insecurity and inferiority of the psychologist who is anxious that his research effort yield a significant contribution, and so to buttress his own security seeks to embark on a second research project simultaneously. But, even this does not assuage his anxiety so that he must turn now to a third project, hoping that in some way a multiplicity of research activities will at least yield one research project which will earn him success. This multiplicity of efforts frequently leads to activity but no sustained research.

The Clinician as a Researcher

A question that frequently is posed is whether all clinical psychologists should be expected to do research. Certainly, at the very least, we should expect that the individual who has received a PhD in clinical psychology approach his clinical tasks with an orientation geared to recognize problems that remain unanswered and questions seeking solutions. Perhaps this is as much as we should expect from the bulk of practicing clinical psychologists—this capacity to recognize and be aware of the problems needing solution. Thus, an intern may conceivably have a successful experience with regard to research if he can spend his year within a setting that, while it does not practice research, at least is self-questioning and self-analyzing with regard to its beliefs and its techniques. However, it is the thesis of the author that the more significant research experience during the internship will be accomplished where the setting is itself one where the psychologists are actively involved in research activity.

The controversy as to whether we should be training people to do research or to evaluate published research still continues. Even if we could agree that this was a legitimate conflict, in what way would the training for doing research be different from the training that would enable one to evaluate research? Is it possible to evaluate research without systematic involvement in research that brings home to one realistically the important issues with which one is concerned in research activity?

It is important to stress that there are many skills that enter into being a well-rounded research person. Among these are: the capacity to recognize meaningful research problems; the capacity to engage in the creative thinking involved in building hypotheses and deducing implications from them; the capacity to integrate and utilize theory to refine a problem; the capacity to develop a research plan with appropriate controls that would make possible the testing of the hypotheses; the capacity to assess the nature of populations and the types of samplings to be made; the capacity to select the appropriate techniques and instruments to be used in the measuring process of the research; the capacity to deal with subjects not as clinical but as research entities; the capacity to record faithfully responses made by subjects; the capacity to tabulate, score, and analyze the data resulting from the research; and the capacity to communicate, both in written and oral form, the nature of the investigation and the results. Clearly, there are few people who possess all of these skills to a sufficient degree to qualify as the well-rounded research clinician. Here individual differences must be recognized and accepted. Individuals vary quantitatively in the extent to which they possess the various skills that have been enumerated, and a place in research exists for all people trained as clinical psychologists, each one making a contribution commensurate with his capacities and within his limitations. Group research is perhaps a prime way in permitting every clinical psychologist a research role.

Research Training for the Intern

It seems imperative that the setting in which the intern receives his first major clinical experience be one in which research in all of its aspects is a vibrant part of the program. This is especially true with regard to thinking about research. Toward this end,

we have found it advisable to have frequent conferences devoted either to ongoing research or published research. Even in those settings that may not be able to integrate internes into ongoing research, it seems advisable that this provision for regular discussions should become institutionalized as part of the program..

One problem that presents itself in many clinical settings is that of providing time for research activity in the face of service demands. We have utilized the technique of releasing the intern every third week from diagnostic responsibilities in order to permit him to work on research. The importance of providing time is that it is a concrete demonstration of the department's and the agency's attitudes toward research. This open acceptance of research activity as a legitimate function of the clinician will minimize the anxiety or guilt in the individual who feels that he is sacrificing services for patients by engaging in research.

Another problem with which we have struggled has been that of organizing interns so that they may participate in group research which provides a richer learning experience for the intern than he can normally expect from individual research. The opportunity for continued conferences on the group research introduces a significant dimension into the training of the intern that can hardly be matched when the intern is working alone on his own problem, even where he is being supervised. Working in the group on a group research does not minimize the opportunity for individual supervision or for contact with consultants, but rather adds to this the involvement with a group of other psychologists with whom the intern can interact on research in a continuous way over a year. Besides the potentially greater intellectual stimulation made possible by group research, there is the support that is provided in dealing with the many complex problems that arise in the course of research activity.

The use of research consultants is felt to be a very valuable part of the research training experience for the intern. Not only do they add to the research atmosphere of the department, but also these people, when carefully selected, can provide genuine help on research problems from the theoretical conceptualization to the technical aspects of designing the research and analyzing the data. The use of these consultants may also help the intern to understand that there may not be as great a cleavage between the university research psychologist and the "field" clinician as many interns seem to feel. By observing the fact that these two can communicate, share, and sometimes collaborate to their mutual benefit, the intern learns that there need not be this sharp schism between clinical activity and research as it is often defined in the university research center.

A problem that must frequently be faced is the latent feeling in some interns that, if they engage in agency research, they will somehow be misused. This misuse may occur at two levels: (*a*) The intern may fear that he will be relegated to tasks that are either demeaning or more appropriate to an individual with less training. This suggests that, in providing a training experience in research for the intern, it must be an experience that will be truly an educational one. He cannot be used simply as an assistant or as a clerical worker, but must share directly in the research from its conceptualization to its execution. (*b*) The intern may fear that he will not be rewarded for his contributions to the research in the form of authorship. In clinical activity, this concern is not present. Whether he is performing as a diagnostician or therapist, the intern is explicitly recognized for his services. Since the diagnostic workup and the record of treatment of the patient in the medical file bear his name, this serves to provide the intern with formal recognition of his labors. In research, this immediate recognition is not readily apparent. It has been our experience that with some interns this is a significant concern that may contribute to the intern's reluctance to become involved in hospital research. We have found that it is well to be explicit with interns with regard to the rewards that they may expect from participation in the research. This has most usually meant that the intern is accepted as a collaborator from the point of view of authorship.

With the provision for the intern to work on a clinical research problem during the in-

ternship, there carries with it a requirement that such research should be supervised by someone who not only can assess the research from a technical point of view (design, sampling, etc.), but can also make the research as clinically meaningful as possible. It is the rare intern who can function independently with regard to research activity, even when he has already completed his doctoral dissertation. We have recently introduced, on a trial basis, a research training program for interns which provides for the joint supervision of interns' research by a university experimentalist and by a clinician.

Part of the responsibility of the research worker is to communicate his research and its findings effectively both in written and in oral form. It is part of the training of the intern that once he is involved in his research he is expected to prepare reports to be presented at group conferences at which consultants may be present.

Conclusions

It may be possible to reconceptualize the problem of how one trains the intern to perform a research role effectively. In a sense, we may say that the individual learns to do research most effectively when he is globally involved in this process of learning. That is, he learns first by feeling that research is an activity personally and professionally rewarding and by feeling comfortable in performing this research role because it does not conflict with other roles or with his own self-concept; he learns by thinking research, by being involved in activities that will permit him to extend his intellectual horizons with regard to research and particularly clinical research; and finally, he learns by doing research, by actually becoming involved in research activity. Our emphasis here has been that the internship's responsibility with regard to developing the research role of the clinician must be to permit the intern to begin to *feel* like a research person, to provide experiences that will extend his *thinking* about clinical problems and methods for dealing with them, and to provide an experience of *doing* in which he may practice the skills of a research worker.

References

Holzberg, J. D. The clinical and scientific methods: Synthesis or antithesis? *J. proj. Tech.*, 1957, **21**, 227–242.

Kubie, L. S. Some unresolved problems of the scientific career. *Amer. Scientist*, 1953, **41**, 596–613.

Redlich, F. C., & Brody, E. B. Emotional problems in interdisciplinary research in psychiatry. *Psychiatry*, 1955, **18**, 233–239.

Sarbin, T. R. Role theory. In G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge: Addison-Wesley, 1954.

(Received April 11, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 192–195

COMPARABILITY OF INTELLIGENCE QUOTIENTS OF MENTAL DEFECTIVES ON THE WECHSLER ADULT INTELLIGENCE SCALE AND THE 1960 REVISION OF THE STANFORD–BINET[1]

GARY M. FISHER,[2] BEVERLY A. KILMAN, AND ANNA M. SHOTWELL
Pacific State Hospital, Pomona, California

A number of studies comparing the Wechsler-Bellevue (W-B) and the Wechsler Intelligence Scale for Children (WISC) with the Stanford-Binet (S-B) have been reported. Correlation coefficients between IQs from these Wechsler scales and the S-B have been of the same order (.6 to .9) in normal, neuropsychiatric, and mentally retarded populations (Alderdice & Butler, 1952; Benton, Weider, & Blauvelt, 1941; Goldfarb, 1944; Gothberg, 1949; Mitchell, 1942; Nale, 1951; Sandercock & Butler, 1952; Stacey & Levin, 1951; Wechsler, 1944). Although some studies have indicated IQs from the WISC to be slightly higher than those from the S-B, the two instruments give fairly comparable scores (Frandsen & Higginson, 1951; Gehman & Matyas, 1956; Harlow, Price, Tatham, & Davidson, 1957; Weider, Noller, & Scramm, 1951). A number of studies have indicated that W-B IQs are consistently higher than S-B IQs of older subjects (Halpern, 1942; Mitchell, 1942) and that this discrepancy becomes larger as age increases and intellectual level decreases (Bensberg & Sloan, 1950; Kutash, 1945; Mundy & Maxwell, 1958).

Only one study, that of Wechsler (1958), compares the Wechsler Adult Intelligence Scale (WAIS) with the S-B. In a sample of 52 male prisoners, aged 16 to 26 years, the mean S-B IQ was five points higher than the mean WAIS IQ, and the correlation between the two IQs was .85. The present study compares adult hospitalized retardates of varying age and intelligence level as to their IQs on the WAIS and the 1960 revision of the Stanford-Binet Form L-M.

METHOD

The sample consisted of 180 mentally retarded subjects in three California state hospitals who were 18 years or older and who had a diagnosis of familial or undifferentiated mental retardation.[3] The sample was classified into four unequal age groups and three unequal IQ categories, in order to obtain 12 equal-sized subgroups of 15 patients each. The age grouping by years was: 18–34, 35–44, 45–54, and 55–73. The grouping by intellectual level (based on the average of WAIS and S-B IQ)[4] was: 46 and below, 47–54, and 55 and over.

Half of the subjects were given the WAIS first and the remainder the S-B first. For a given subject, different examiners administered the two tests. A counterbalancing procedure was employed in assigning subjects to the two testers in order to control for possible examiner difference. To control for changes

[1] This study was supported by a grant from the California Department of Mental Hygiene (No. R-58-14.1-3). The authors gratefully acknowledge the assistance of Edward Bartsch, Milton Dooley, James Judge, Robert Kirby, Earl Owens, Dorian Rose, and Charles Windle. Special thanks are due Maud Merrill for supplying material for the 1960 Revision of the Stanford-Binet before its publication (Terman & Merrill, 1960) and to Arthur Silverstein for his assistance with the statistical analyses.

[2] Now at Fairview State Hospital, Costa Mesa, California.

[3] Thus subjects with known brain damage were excluded from the study.

On the WAIS, a subject must achieve a minimal Total Scaled score of 11 to obtain an IQ from the table of norms. Consequently, any subject achieving less than this minimum score was excluded from the sample since only prorated IQs could be estimated and presumably the WAIS would not be an appropriate test for such a subject.

[4] This procedure for determining IQ level was adopted in order to circumvent the problem of statistical regression toward the mean which would occur if IQ level were determined by only one test.

TABLE 1

COMPARISON OF WECHSLER ADULT INTELLIGENCE SCALE AND STANFORD-BINET IQs FOR EACH AGE LEVEL

Age	WAIS IQ			S-B IQ			*t* value	Correlation WAIS & S-B
	Mean	*SD*	Range	Mean	*SD*	Range		
18–34	58.42	8.40	45–77	44.02	9.25	26–68	15.00	.736
35–44	59.44	10.20	45–96	45.84	10.99	31–75	12.04	.752
45–54	60.29	8.31	48–76	43.51	8.61	26–68	18.64	.752
55–73	63.56	9.88	51–87	41.00	8.33	26–57	23.92	.777

which might occur in intellectual functioning over an extended period of time, all subjects were given both tests within a 12-month period, with the great majority of subjects receiving both tests within a 4-month period.

To secure data that might help in developing an hypothesis regarding the relation of social competency to intellectual functioning on the WAIS and S-B, all subjects were administered the Vineland Scale of Social Maturity (Doll, 1953). In addition, an adapted form of a Scale of Minimal Social Behavior (Farina, Arenberg, & Guskin, 1957) was administered, but because of the lack of variability in scores from this instrument no use could be made of the results.

RESULTS

An analysis of variance of the difference scores between WAIS and S-B IQs was performed. The results indicated that age was significant in determining the magnitude of the discrepancy between the two IQs ($F = 16.74$; $df = 3$; $p < .001$), but neither IQ level nor the interaction between IQ level and age was significant.[5]

Table 1 shows the comparison of WAIS and S-B IQs for each age level. Of the 180 subjects examined only 3 had an S-B IQ higher than their WAIS IQ, the largest difference being five points. The *t* values for the significance of the difference between correlated means indicated that the mean WAIS IQ was significantly larger (beyond the .001 level) than the S-B IQ at each age level. The correlation coefficients between WAIS and S-B IQs ranged from .736 to .777 and a chi square test (Edwards, 1950) indicated that the differences among them were not significant ($\chi^2 = .20$; $df = 3$; $p > .95$).

Table 2 sets forth the difference scores between WAIS IQ and S-B IQ by age level. The *t* tests indicated that the mean difference scores for the three younger age groups were significantly smaller ($< .01$) than for the oldest age group. Between the ages of 18 and 54 years, the WAIS IQ averaged 15 points higher than the S-B IQ, whereas in subjects over 55 years the difference averaged 23 points. The standard deviation of the difference score was approximately 6. The WAIS, unlike the S-B, takes systematic account of the lower performance of older adults; hence, the older the adults (beyond a peak between 25 and 30) the lower the absolute level of performance and hence the larger the discrepancy between S-B and WAIS. As the WAIS standardization population included adults of a wide age range whereas the S-B did not, it would be logical to assume that the WAIS IQ is a more accurate measure of in-

TABLE 2

DIFFERENCE SCORES BETWEEN WECHSLER ADULT INTELLIGENCE SCALE IQs AND STANDARD-BINET IQs

Age	Difference Scores		
	Mean	*SD*	Range
18–34	14.49	6.49	−5 to 32
35–44	13.60	7.53	−5 to 27
45–54	16.56	5.75	5 to 32
55–73	22.56	6.28	10 to 36

Note.—For each subject, the S-B IQ was subtracted from the WAIS IQ.

[5] Analyses of the difference scores between WAIS Verbal IQ and S-B IQ, and between WAIS Performance IQ and S-B IQ were also performed. The results were virtually identical with the analysis of the difference scores between WAIS and S-B IQs. Moreover, these two sets of difference scores were of the same magnitude as difference scores between WAIS Full Scale IQs and S-B IQs.

TABLE 3

COMPARISON OF IQs FROM THE WECHSLER ADULT INTELLIGENCE SCALE AND STANFORD-BINET WITH SOCIAL AGES FROM THE VINELAND SCALE OF SOCIAL MATURITY

Age	WAIS Mean IQ	S-B Mean IQ	Vineland Scale of Social Maturity			Correlation Coefficients	
			Mean Age	*SD*	Range	WAIS IQ & Social Age	S-B IQ & Social Age
18–34	58.42	44.02	10–2	1–6	5–6 to 13–3	.265	.322
35–44	59.44	45.84	10–2	1–9	5–5 to 13–0	.438	.171
45–54	60.29	43.51	9–9	2–0	4–0 to 12–10	.638	.433
55–73	63.56	41.00	9–7	2–1	5–5 to 13–3	.534	.670

telligence. It has been suggested, however, that there is a spurious allowance for "normal" deterioration with age for retardates on the Wechsler scales (Bensberg & Sloan, 1950). Until validity studies, preferably of the longitudinal type, of the WAIS with retarded populations are made, the problem of differential rate of intellectual decline depending on level of intellectual functioning remains unanswered.

The regression equations for predicting the IQ on one test from that on the other for the three younger age groups appeared sufficiently similar to warrant the use of the following regression equations for subjects between 18 and 54 years:

Predicting WAIS IQ from S-B IQ:

$$Y = .68X + 29.15 \qquad (SE_e = 6.15)$$

Predicting S-B IQ from WAIS IQ:

$$Y = .79X - 2.45 \qquad (SE_e = 6.61)$$

For subjects 55 years and older, the following regression equation should be used:

Predicting WAIS IQ from S-B IQ:

$$Y = .92X + 25.84 \qquad (SE_e = 6.22)$$

Predicting S-B IQ from WAIS IQ:

$$Y = .65X - .31 \qquad (SE_e = 5.25)$$

In order to gain some knowledge of the relation between social competency and the IQs from the two intelligence scales, comparisons of the Social Ages (SA) from the Vineland Scale of Social Maturity were made with the IQs from the WAIS and the S-B. These data are presented in Table 3. Chi square tests indicated that the correlation coefficients between WAIS IQ and Vineland SA and between S-B IQ and Vineland SA were homogeneous for the four age levels ($\chi^2 = 5.249$ and 9.264, respectively; $df = 3$; $p > .01$). For the total sample the correlation coefficient between Vineland SA and WAIS IQ was .450, and between Vineland SA and S-B IQ was .405. A nonsignificant t value of .939 for the significance of the difference between correlated correlations indicated that the WAIS and S-B correlated equally well with Vineland SA. The data shown in Table 3 suggest a trend for WAIS IQs to increase, and S-B IQs and Vineland SA to decrease, with age. According to analyses of variance these observed trends all proved to be nonsignificant at the .01 level (WAIS: $F = 2.54$; S-B: $F = 2.01$; Vineland SA: $F = 1.29$). More thorough investigation would be necessary in order to answer the question of the relation between social competency and intelligence as measured by these two scales.

SUMMARY

This study sought to determine the effect of age and level of retardation on the comparability of IQs from the Wechsler Adult Intelligence Scale and the 1960 revision of the

Stanford-Binet. In addition, a measure of social competency was related to the IQs from the two scales. It was determined that age, but not level of retardation, was significant in determining the magnitude of the difference between WAIS and S-B IQs. WAIS IQs averaged 15 and 23 points higher than S-B IQs for subjects 18–54 years and 55–73 years, respectively. Regression equations were calculated to translate the IQ from one test to the other test. The WAIS and S-B IQs correlated equally well with Social Ages from the Vineland Scale of Social Maturity.

REFERENCES

ALDERDICE, E. T., & BUTLER, A. J. An analysis of the performance of mental defectives on the Revised Stanford-Binet and the Wechsler-Bellevue. *Amer. J. ment. Defic.*, 1952, **56**, 609–614.

BENSBERG, G. J., & SLOAN, W. Wechsler's concept of "normal deterioration" in older mental defectives. *J. clin. Psychol.*, 1950, **6**, 359–362.

BENTON, A. L., WEIDER, A., & BLAUVELT, J. Performance of adult patients on the Bellevue intelligence scales and the revised Stanford-Binet. *Psychiat. Quart.*, 1941, **15**, 802–806.

DOLL, E. A. *Measurement of social competence*. Minneapolis: Educational Publishers, 1953.

EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.

FARINA, A., ARENBERG, D., & GUSKIN, S. A scale for measuring minimal social behavior. *J. consult. Psychol.*, 1957, **21**, 265–268.

FRANDSEN, A. N., & HIGGINSON, J. B. The Stanford-Binet and Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 236–238.

GEHMAN, ILA H., & MATYAS, R. P. Stability of the WISC and Binet tests. *J. consult. Psychol.*, 1956, **20**, 150–152.

GOLDFARB, W. Adolescent performance in the Wechsler-Bellevue and the revised Stanford-Binet examination, Form L. *J. educ. Psychol.*, 1944, **35**, 503–507.

GOTHBERG, LURA C. A comparative study of the Stanford-Binet old form test and the Wechsler-Bellevue verbal, performance, and full scale as shown in the results of unselected employees. *Amer. J. ment. Defic.*, 1949, **53**, 497–503.

HALPERN, FLORENCE. Comparison of the revised Stanford-Binet with the Bellevue adult intelligence scale. *Psychiat. Quart. Suppl.*, 1942, **16**, 206–211.

HARLOW, J. E., PRICE, A. C., TATHAM, L. J., & DAVIDSON, J. F. Preliminary study of comparison between Wechsler Intelligence Scale for Children and Form L of revised Stanford-Binet scale at three age levels. *J. clin. Psychol.*, 1957, **13**, 72–73.

KUTASH, S. B. A comparison of the Wechsler-Bellevue and the revised Stanford-Binet scales for adult defective delinquents. *Psychiat. Quart. Suppl.*, 1945, **19**, 677–685.

MITCHELL, M. B. Performance of mental hospital patients on the Wechsler-Bellevue and revised Stanford-Binet Form L. *J. educ. Psychol.*, 1942, **33**, 538–544.

MUNDY, LYDIA, & MAXWELL, A. E. The assessment of the feebleminded. *Brit. J. med. Psychol.*, 1958, **31**, 201–211.

NALE, S. The children's Wechsler and the Binet on 104 mental defectives at the Polk State School. *Amer. J. ment. Defic.*, 1951, **56**, 419–423.

SANDERCOCK, MARIAN G., & BUTLER, A. J. An analysis of the performance of mental defectives on the Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1952, **57**, 100–105.

STACEY, C. L., & LEVIN, JANICE. Correlation analysis of scores of subnormal subjects on the Stanford-Binet and Wechsler Intelligence Scale for Children. *Amer. J. ment. Defic.*, 1951, **55**, 590–597.

TERMAN, L. M., & MERRILL, MAUD A. *Stanford-Binet intelligence scale*. Boston: Houghton Mifflin, 1960.

WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.

WECHSLER, D. *The measurement and appraisal of adult intelligence*. (4th ed.) Baltimore: Williams & Wilkins, 1958.

WEIDER, A., NOLLER, P. A., & SCRAMM, T. A. The WISC and the Stanford-Binet. *J. consult. Psychol.*, 1951, **15**, 330–333.

(Received April 1, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 196–199

PERCEPTUAL SIZE CONSTANCY IN CHRONIC SCHIZOPHRENIA[1]

H. W. LEIBOWITZ
University of Wisconsin
AND VLADIMIR PISHKIN
Veterans Administration Hospital, Tomah, Wisconsin

The ability to judge correctly the sizes of objects despite variation in viewing distance, i.e., size constancy, represents an important biological achievement of living organisms and has been studied in relation to the underlying mechanisms as well as to psychopathology. In particular, the possibility that size constancy differs among schizophrenics as compared with normals has been of interest to a number of investigators and theorists (Lovinger, 1956; Maes, 1957; Raush, 1952, 1956; Reynolds, 1954; Sanders & Pacht, 1952; Weckowicz, 1957). Bruner (1951) states the theoretical problem most succinctly when he suggests that a withdrawal from object relations and an increasing concern for the self might lead to a breakdown in perceptual constancy. However, as pointed out in a recent summary (Rabin & King, 1958), there is little agreement among the results of previous experiments. Sanders and Pacht (1952) found "overconstancy" among a schizophrenic group as compared with normals, while Weckowicz (1957) reports the opposite relationship. Raush (1956) found no differences between nonparanoid schizophrenics, although his paranoid group did differ from normals.

The purpose of the present experiment is to re-examine this problem, utilizing a group of chronic, undifferentiated schizophrenics chosen from the patient population of a neuropsychiatric hospital to be most "withdrawn" with respect to their behavior. It is felt that the selection of the extreme cases from among a group of chronic patients whose diagnosis is more reliable provides an excellent test of the hypothesis under consideration. The method of testing has been modified after techniques which have been employed in experimental laboratories and for which a large body of normative data is already available (Holway & Boring, 1941; Leibowitz, in press; Leibowitz, Chinetti, & Sidowski, 1956). Essentially, the technique requires the subject to signal which of two sticks is larger, a simple task which has been used successfully with children and feebleminded groups. Data are obtained over a wide range of viewing distances, thus providing the basis for a more complete analysis of the size matching–distance relationship.

[1] Supported, in part, by Grant M-1090 from the National Institute of Mental Health, National Institutes of Health, United States Public Health Service, and the Veterans Administration Research Program.

METHOD

Subjects

The members of the experimental group were 35 male patients of the Veterans Administration Hospital in Tomah, Wisconsin, carrying a diagnosis of chronic, undifferentiated schizophrenia. There was no history of brain damage, organic pathology, or visual defect. None of the patients were receiving EST, although many of them were on different types of drug therapy. They were selected by the psychiatric ward team to be mainly characterized by withdrawal symptoms. The mean age of the 35 patients was 39.41 years with a range from 24 years to 56 years. The average length of psychiatric hospitalization was 8.84 years with a range of from 2 years to 13 years.

The members of the control group were 20 psychiatric aides and were selected at random. Their mean age was 38.52 years with a range of from 25 years to 54 years.

Experimental Conditions

The experiment was conducted in a hospital corridor 9½ feet wide, 9½ feet high, and 135 feet in

length. Large windows were located every 12 feet in both walls of the corridor and all tests were made in daylight. The test objects consisted of five wooden dowels cut from 1 inch stock to lengths of 2, 4, 8, 16, and 24 inches. These were mounted in square bases 4 inches by 4 inches and were painted "flat" black. The viewing distance of the test objects was proportional to their size, the ranges being 10, 20, 40, 80, and 120 feet, respectively. Thus, the visual angle subtended by each object, and the size of the corresponding retinal image, was the same in each case (0.96 degree of arc). The subject was seated at one end of the corridor and first shown one of the test objects set up at its appropriate viewing distance. The subject was told to look at the height of the stick and to tell whether a comparison stick, which was placed at right angles to the line of sight to the test object and 8 feet from the subject, was shorter or taller than the test object. The comparison objects were chosen from a series ranging in size from 1 inch to 35 inches and were indistinguishable, except for size, from the test objects. The order of presentation of the comparison stimuli was chosen so that the first comparison would easily elicit a "higher" or "shorter than" response from the subject, while the second reversed his response. Subsequent sticks were chosen to be progressively nearer the point of subjective equality. The order of presentation of the comparison objects was determined by a 5×5 latin square design. No difficulty was encountered in testing the patients. They cooperated well, exhibited no difficulty in understanding the instructions nor in making judgments, and showed no signs of complaint throughout the study, which lasted approximately 15 minutes for each patient. The two authors served alternately as experimenters.

Results

The mean sizes of the comparison objects which matched the test objects at the various distances employed are given for both groups along with their standard deviations in Table 1. Figure 1 represents a plot of mean matched size as a function of test object distance. On this figure a horizontal line represents a prediction in terms of the "law of visual angle,"

TABLE 1

Matched Size of a Series of Test Objects as a Function of Distance for Schizophrenic and Normal Subjects

Test Object Size (Inches)	Test Object Distance (Feet)	Schizophrenic ($N = 35$)		Normal ($N = 20$)	
		Mean	*SD*	Mean	*SD*
2	10	2.08	0.23	2.13	0.22
4	20	4.11	0.19	4.11	0.36
8	40	9.09	1.09	8.97	1.08
16	80	17.87	1.74	17.03	1.88
24	120	24.71	2.99	25.48	2.41

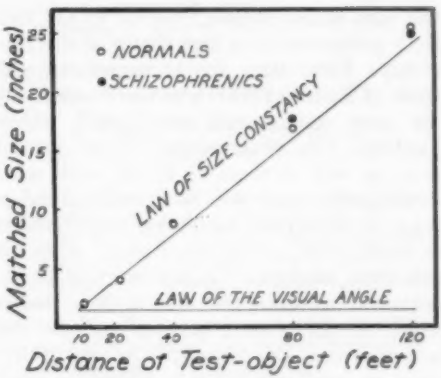

Fig. 1. Matched size as a function of distance for a group of normals and of chronic schizophrenics. (The test objects subtended the same visual angle at all distances of observation.)

a condition which would obtain if judged size depended only on retinal image size. In the present study, retinal image size was constant so that matched size, according to this prediction, would be the same for all test objects and viewing distances. The other theoretical extreme, the "law of size constancy," is represented by the diagonal line. According to this prediction, perceived size is independent of distance so that matched size would be equal to the actual size of the test objects. As is typical in such experiments conducted with binocular vision, adequate illuminance, and a well articulated visual field, the data for normals lie very close to the prediction in terms of size constancy. The data for the schizophrenics also lie close to this theoretical line.

The nature of the data indicated the use of a nonparametric test. The Mann-Whitney U test was applied (Siegel, 1956). From this analysis it was concluded that size judgments of the schizophrenic group do not differ significantly from the normal sample on this task ($U = 119$, $p < .424$). In addition, respective variabilities of the two groups were compared. As expected, none of the differences reached significance.

Discussion

The results of the present study indicate no difference between the ability of chronic schizophrenics and normals to judge correctly

the sizes of test objects from 10 to 120 feet. Both groups exhibit a high degree of size constancy. These data are in agreement with those of Raush (1952), who found essentially the same results with nonparanoid schizophrenics. The disagreement of the present data, as well as those of Raush, with other investigators could well be a result of differences in the patient population (e.g., Sanders & Pacht, 1952, tested outpatients) or in the procedure employed. In any event, it is important to recognize that even in those studies which do differentiate between normals and schizophrenics the magnitude of the differences are small in relation to differences found as a result of variation of less subtle variables, such as the instructions given the subjects (Gilinsky, 1955; Holaday, 1933), or the age of the subjects (Beryl, 1926; Zeigler & Leibowitz, 1957).

The present data have implications in relation to psychopathology, as well as to a theoretical understanding of size constancy. With respect to the chronic schizophrenic group tested, it is clear that whatever disorders of perception, thinking, or behavior they suffer from, their size constancy is unaffected. Indeed, it would be difficult to imagine such patients, who move about the hospital grounds and engage in sports, judging incorrectly the sizes of environmental objects. The difficulty with schizophrenics, as seen clinically, may not be in relation to neutral objects, such as the sticks utilized in this study, but to other individuals and to their own patterns of thinking and emotions. Apparently, the withdrawal characteristic of schizophrenia, in general, and of the subjects utilized in this study, in particular, may be significant in interpersonal relationships only, and are not indiscriminately employed as defenses in purely nonaffective areas, such as size judgment. As suggested by Bruner (1951), one would be more likely to find differences between schizophrenics and normals if the "test objects" were human beings.

The present size matching data are also of importance in relation to the mechanisms which subserve the size constancy effect. It has been demonstrated that the size constancy of children for near objects is good, but that the ability to judge correctly the sizes of distant objects develops slowly (Zeigler & Leibowitz, 1957). These data suggest that there are a number of separate mechanisms involved in size constancy, the nearer distances being mediated by the kinesthetic cues of accommodation and convergence, the farther by a more slowly developing perceptual learning process (Leibowitz & Moore, 1960). The available data indicate that this learning process is complete by the early teens or, in any event, since the patients in the present study were veterans, before the age at which the subjects tested were hospitalized. The lack of any difference between the groups in the present experiment suggests that the mechanisms underlying size constancy for neutral objects, such as used in this study, are independent of personality changes even as severe as those encountered in schizophrenia. This conclusion is in agreement with similar studies which demonstrate that the development of size constancy is independent of intellectual processes as indicated by the lack of differences between feebleminded and normal subjects on size matching tasks (Jenkin & Morse, in press; Leibowitz, in press). Thus, it would appear that the development of size constancy, although closely related to the age of the subject, is independent of mental and personality development.

Summary

The ability to judge object size as a function of distance was determined for a group composed of chronic, undifferentiated schizophrenics, as well as a control group of psychiatric aides.

There were no significant differences in the matches produced by the two groups. Both judged correctly the sizes of the test objects at all distances.

It is suggested that the absence of any differences is due to the fact that size matching requires abilities which are fully developed prior to the onset of schizophrenia and which are unaffected by the characteristic withdrawal observed in this pathology.

REFERENCES

BERYL, F. Über die Grössenauffassung bei Kindern. *Z. Psychol.*, 1926, **100**, 344–371.

BRUNER, J. S. Personality dynamics and perceiving. In R. Blake & G. Ramsey (Eds.), *Perception: An approach to personality*. New York: Ronald, 1951. Pp. 121–145.

GILINSKY, A. The effect of attitude upon the perception of size. *Amer. J. Psychol.*, 1955, **68**, 173–192.

HOLADAY, B. E. Die Grössenkonstanz der Sehdinge bei Variation der inneren und äusseren Wahrnehmungsbedingungen. *Arch. ges. Psychol.*, 1933, **88**, 419–486.

HOLWAY, A. H., & BORING, E. G. Determinants of apparent visual size with distance variant. *Amer. J. Psychol.*, 1941, **54**, 21–37.

JENKIN, N., & MORSE, S. Developmental and intellectual processes in size-distance judgment. *Amer. J. Psychol.*, in press.

LEIBOWITZ, H. Apparent visual size as a function of distance for mentally deficient subjects. *Amer. J. Psychol.*, in press.

LEIBOWITZ, H., CHINETTI, P., & SIDOWSKI, J. Exposure duration as a variable in perceptual constancy. *Science*, 1956, **123**, 668–669.

LEIBOWITZ, H., & MOORE, D. The role of oculomotor adjustments in the perception of size. *J. Opt. Soc. Amer.*, 1960, **50**, 507. (Abstract)

LOVINGER, E. Perceptual contact with reality in schizophrenia. *J. abnorm. soc. Psychol.*, 1956, **52**, 87–91.

MAES, J. L. Size constancy in schizophrenia. Unpublished master's thesis, Michigan State University, 1957.

RABIN, A. I., & KING, G. F. Psychological studies. In L. Bellak (Ed.), *Schizophrenia*. New York: Logan, 1958.

RAUSH, H. L. Perceptual constancy in schizophrenia: I. Size constancy. *J. Pers.*, 1952, **21**, 176–187.

RAUSH, H. L. Object constancy in schizophrenia: The enhancement of symbolic objects and conceptual stability. *J. abnorm. soc. Psychol.*, 1956, **52**, 231–234.

REYNOLDS, G. A. Perceptual constancy in schizophrenics and "normals." Unpublished doctoral dissertation, Purdue University, 1954.

SANDERS, R., & PACHT, A. R. Perceptual size constancy of known clinical groups. *J. consult. Psychol.*, 1952, **16**, 440–444.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

WECKOWICZ, T. E. Size constancy in schizophrenic patients. *J. ment. Sci.*, 1957, **103**, 475–486.

ZEIGLER, H. P., & LEIBOWITZ, H. Apparent visual size as a function of distance for children and adults. *Amer. J. Psychol.*, 1957, **70**, 106–109.

(Received April 11, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 200–204

THE RELATIONSHIPS BETWEEN INDIVIDUALLY DEFINED AND GROUP DEFINED SOCIAL DESIRABILITY AND PERFORMANCE ON THE EDWARDS PERSONAL PREFERENCE SCHEDULE

ALFRED B. HEILBRUN, JR. AND LEONARD D. GOODSTEIN
State University of Iowa

Since its publication in 1954, the Edwards Personal Preference Schedule (EPPS) has been the focus of considerable research bearing upon the problem of social desirability as a source of response variance. The EPPS was constructed so that each item consists of a pair of statements matched for social desirability, thus presumably forcing the test taker to respond in accord with his actual behavioral characteristics rather than the social desirability of the statements. To the extent that Edwards (1959b) is correct in considering social desirability as contributing principally to measurement error, studies of the relationship between social desirability and EPPS performance are important in an overall assessment of the test's validity. However, the many investigations have provided rather inconsistent results and conclusions.

Several studies have concluded that social desirability has been eliminated as an important determinant of performance on the EPPS. Navran and Stauffacher (1954) found a near-zero correlation between social desirability ratings of the traits being measured and scores on the EPPS scales. Borislow (1958) reported that subjects were able to change their test profiles under both social and personal desirability faking sets but concluded that desirability had been eliminated as a source of conscious response manipulation on the EPPS because no consistent scale patterns among the subjects were obtained under either set. Finding only a "slightly greater than chance" number of significant point biserial correlations between the subjects' Social Desirability scale scores on the MMPI and frequency of choice of the first or second statement in each item pair of the EPPS led Kelleher (1958) to conclude that social desirability plays an insignificant role in EPPS performance.

In contrast to these negative findings, several investigators have reported results which indicate that desirability remains an important source of response variance on the EPPS. Heilbrun (1958) found a .60 correlation between EPPS scale scores and *personal* desirability of the test scales. Corah, Feldman, Cohen, Gruen, Meadow, and Ringwall (1958) obtained a .88 correlation between the percentage of college subjects judging the first statement in an item as more socially desirable and the percentage of similar subjects in an independent group who endorsed that statement as self-characteristic. Edwards, Wright, and Lunneborg (1959), pointing out that Corah et al. used only 30 EPPS items taken out of context, have repeated the procedure using all items in the test. They report somewhat lower correlations (.69 and .61) for two samples, but the relationship between desirability and endorsement still remained.

One common characteristic of all of these studies described above is that they have sought to evaluate the relationship between desirability and EPPS performance using some type of *group* desirability statistic (e.g., group average or group percentage) in the analysis. It is possible that if the relationship between each subject's *individual* desirability ratings of the statement alternatives in the EPPS and his actual statement selections were determined, a more accurate inference could

be made regarding the influence of social desirability. The logical expectation would be that if social desirability is related to EPPS performance, this should be more evident using an individualized approach, since the individual desirability values which each subject assigns to test statements should be more relevant in predicting *his* statement endorsements than would averaged group values. A recent study by Taylor (1959), however, found the opposite to be the case. He determined the correlations between both individual and group social desirability values and endorsement of MMPI items for schizophrenic patients and found the group-averaged values correlated considerably higher ($r = .79$) with performance than did the individual values ($r = .36$).

The present study had three purposes: (*a*) to relate individual social desirability values and EPPS performance to clarify the extent to which individual social desirability set contributes to response selection; (*b*) to compare the prediction of EPPS responses from individual social desirability values with prediction of these responses from group social desirability values; and (*c*) if individual values are found to be related to overall EPPS performance, to evaluate the specific relationship between individual desirability and each EPPS scale.

Method

Subjects. The 58 subjects used in this study were obtained from a large undergraduate psychology course at the State University of Iowa. This sample included 29 males and 29 females.

Test. The EPPS (Edwards, 1959a) is an objective, rationally derived, multivariate, personality questionnaire. It includes 135 different statements bearing upon personality characteristics, 9 for each of the 15 traits measured by the test. These statements are arranged in 210 pairs with each pair including 2 statements matched for social desirability but measuring different traits. The final score for each trait represents the number of occasions out of a possible 28 on which the subject has selected that trait statement as more characteristic than the paired statement.

Procedure. All subjects were initially group administered the EPPS under standard instructions. About 2 months later the same subjects were given the 135 different statements included in the test and asked to rate each as a trait in other people on a nine-point scale from highly socially undesirable to highly socially desirable. This rating procedure was identical to that described by Edwards (1957) in his initial derivation of statement desirability values. Having obtained the subject's EPPS responses under standard conditions and individual desirability ratings from the nine-point scale for both statements in each item pair, it was possible to determine the percentage of times the subject endorsed as self-characteristic the statement having the higher individual social desirability value *for him*. Items for which the subject had ascribed equal individual social desirability values to both statements were omitted from the analysis.

The second step was to determine the percentage of times each subject endorsed as self-characteristic the more highly socially desirable statement in each item pair where the desirability values were those assigned to the statements by Edwards in his initial construction of the test. This percentage reflects the correspondence between EPPS response and group social desirability values obtained from an independent group of college students. There are 204 of the 210 EPPS items for which the group values of the paired statements differ.

Finally, the values assigned to the nine statements measuring each of the 15 EPPS variables were averaged for each subject, and these mean values defined the subject's individual social desirability ratings for the 15 traits. High and low social desirability groups were then constituted independently for each trait and the mean scale scores for these high and low groups were compared to assess whether the relationship between individual desirability and EPPS performance varied over scales. Since there are established sex differences on the EPPS scales (Edwards, 1959a), precautions were taken to insure approximately equal numbers of males and females in both desirability groups for each scale. This was accomplished by distributing individual social desirability scores for each scale separately by sex and defining high and low groups by cutting at or near the median for both sex distributions. The two high and two low desirability groups for each scale were then recombined into one high and one low group with nearly equal numbers of males and females.

Results

Individual social desirability values and EPPS performance. There was a mean of 150.97 items out of a possible 210 in which the paired statements had discrepant individual social desirability values for the 58 subjects in this study. For those items within which the statement values varied, the subjects averaged endorsing the individually more desirable statement as more self-characteristic 67.16% of the time. This percentage value differs significantly from a 50% chance expectancy ($t = 10.34$ for 57 df, $p < .001$) and clearly indicates that performance on the EPPS is related to the individual ratings of

TABLE 1

COMPARISON BETWEEN EPPS SCALE SCORES FOR THE 15 SEPARATE HIGH AND LOW SOCIAL DESIRABILITY GROUPS

EPPS Scale	High Social Desirability			Low Social Desirability			t[a]
	Mean	*SD*	*N*	Mean	*SD*	*N*	
Achievement	15.69	5.55	29	13.52	4.76	29	1.58
Deference	11.62	4.00	29	10.90	3.04	29	.77
Order	11.86	3.95	29	8.83	3.58	29	3.03**
Exhibition	15.93	2.27	29	14.14	4.58	29	—[b]
Autonomy	15.33	4.04	30	11.89	3.84	28	1.46
Affiliation	17.17	3.25	29	14.72	3.87	29	2.58**
Intraception	20.47	4.68	30	15.71	5.32	28	3.61**
Succorance	11.20	4.69	30	10.71	3.68	28	.44
Dominance	16.86	3.60	29	12.97	4.82	29	3.44**
Abasement	12.63	4.72	30	12.46	5.36	28	.13
Nurturance	16.97	4.75	31	12.93	4.61	27	3.26**
Change	17.39	4.65	28	15.43	5.14	30	1.51
Endurance	14.26	4.16	27	11.23	4.31	31	2.71**
Heterosexuality	18.24	4.94	34	15.00	5.01	24	2.36*
Aggression	14.36	4.62	28	9.50	4.47	30	4.05**

[a] One-tailed t tests were used.
[b] The Cochran-Cox approximation was used because of heterogeneity of variance.
* Significant at .05 level.
** Significant at .01 level.

statement social desirability. For the 58 subjects the range of endorsement percentage in the individually socially desirable direction was from 35.03 to 79.66, the subject providing the lower end of the range being unique in that he was the only subject to fall below the 50% mark.

Individual vs. group social desirability values. When the group mean social desirability values of the statement alternatives were considered, it was found that the subjects endorsed the group-defined higher socially desirable statement on 55.80% of the 204 imperfectly matched items. This result is very similar to that reported by Goodstein and Heilbrun (1959) who found, for an independent sample of 248 college subjects, that the group-defined more socially desirable statement was endorsed 55.92% of the time on the average. The 55.80% value found in the present study differs significantly ($t = 8.99$ for 57 df, $p < .001$) from a chance expectancy of 50%, clearly indicating that performance on the EPPS is also related to group-defined desirability of the statements.

A comparison between the predictableness of EPPS statement endorsement from group social desirability values and from individual social desirability values was made, and it was found that the difference between the mean percentage of 67.16 for items in which the endorsed statement was the more individually desirable was significantly higher than the 55.80% of items in which the endorsed statement was group-defined more desirable ($t = 9.86$ for 56 df, $p < .001$).

Individual social desirability and separate EPPS scales. Although it had been shown in this study that the individual's judged desirability of the statement alternatives does show a significant relationship to statement endorsement on the EPPS, the question still remained as to the extent of this relationship for the separate scales of the test. Data pertinent to this question are presented in Table 1 which gives the means and standard deviations on each of the 15 EPPS scales for the 15 separate high and low social desirability groups and the results of t test comparisons. It can be seen that the group of subjects rating the trait statements as more highly desirable had the higher mean on each scale when compared to the group rating the trait statements as less desirable, the differences being significant for 9 of the 15 comparisons and approaching significance at less than the .10

level on 3 more. Thus the positive relationship between individual desirability and endorsement appears to hold for most but not for all of the EPPS scales.

Discussion

One major finding in the present study is that college subjects endorsed the individually defined, more socially desirable response as self-characteristic on at least two out of every three items in which the social desirability values of the paired statements differed. This would seem to clearly indicate that social desirability has not been eliminated as an important source of performance variance on the EPPS as has been suggested by some previous investigators (Borislow, 1958; Kelleher, 1958; Navran & Stauffacher, 1954). The inconsistencies in results among the many investigators bearing upon this issue most likely has stemmed from the use of group social desirability statistics which, based on the results of the present study, are less likely to be related to the individual's response endorsement than the desirability values for that particular individual.

A second major interest in this study was whether an individual's responses to a structured personality questionnaire are more highly related to his own judged social desirability of the responses or to averaged group desirability values. Taylor's (1959) recent finding that group values were more highly related appeared to be inconsistent with the logical assumption that the closer individual parameters of behavior are approximated, the more effectively can behavior be predicted for the individual. The results of the present study were not in agreement with those of Taylor, since it was found that college subjects selected the statement alternative on the EPPS which was individually more desirable more than 67% of the time while these same subjects endorsed the group-defined more desirable statement on less than 56% of the items. Taylor suggested that the lesser relationship obtained between individual social desirability values and MMPI performance than between group values and performance could be attributed in large measure to the lower reliability of individual values which, in turn, would serve to more seriously attenuate this correlation. If this is true, it should be pointed out that the opposite results were found in the present study *despite* the lower reliabilities of the individual social desirability values. Since there were so many procedural differences between the Taylor study and the present one (e.g., nature of test item content, true-false vs. forced choice format of the tests, nature of the subject population, etc.,) further speculation regarding the differential results appears unwarranted at the present time.

The analysis of relationships between individual desirability and specific scores suggests a fairly general positive relationship over scales. An interesting finding was that the three scales which appear to be unrelated to social desirability (Deference, Succorance, and Abasement) have a common psychological element: their characteristic behaviors involve some type of subordination of the individual to another person. This suggests the hypothesis that the young adult college subjects used in this study, being in transition from a period of subordinate childhood relationships into a period of more superordinate or coequal adult relationships, have not as yet stabilized their value systems regarding subordinate, equal, and superordinate interpersonal roles. A consequence of this would be that what the subject perceived as socially desirable and self-characteristic relative to role-related behaviors would be at least temporarily unrelated.

Summary

This study investigated three issues relative to performance on the Edwards Personal Preference Schedule: (*a*) the relationship between the social desirability of statement alternatives on the questionnaire and the endorsement of statements as self-characteristic when the individual's own desirability values are used as predictors, (*b*) the prediction of statement endorsement using the individual's social desirability values for the statements as opposed to prediction using group-defined values for the statements, (*c*) the differential relationships between the individual's social desirability values and performances on the separate scales of the EPPS.

The results indicated that the individual's social desirability set *is* an important source of variance in EPPS performance, since college subjects endorsed as self-characteristic the more highly valued statement at least two out of every three occasions when the statement alternatives were assigned different individual social desirability values. It was also found that individual social desirability values were more highly related to EPPS performance than were group desirability values, a finding which differs from that reported by Taylor (1959) who used the MMPI as the performance variable. Finally, analysis by separate scales suggested that individual social desirability set is related to most but not all of the EPPS variables.

REFERENCES

Borislow, B. The Edwards Personal Preference Schedule and fakability. *J. appl. Psychol.*, 1958, **42**, 22–27.

Corah, N. L., Feldman, M. J., Cohen, I. S., Gruen, W., Meadow, A., & Ringwall, E. A. Social desirability as a variable in the Edwards Personal Preference Schedule. *J. consult. Psychol.*, 1958, **22**, 70–72.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Edwards, A. L. *Personal Preference Schedule.* New York: Psychological Corp., 1959. (a)

Edwards, A. L. Social desirability and personality test construction. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment.* Princeton: Van Nostrand, 1959. (b)

Edwards, A. L., Wright, C. E., & Lunneborg, C. E. A note on "Social desirability as a variable in the Edwards Personal Preference Schedule." *J. consult. Psychol.*, 1959, **23**, 558.

Goodstein, L. D., & Heilbrun, A. B. The relationship between personal and social desirability scale values of the Edwards Personal Preference Schedule. *J. consult. Psychol.*, 1959, **23**, 183.

Heilbrun, A. B. Relationships between the Adjective Check-List, Personal Preference Schedule, and desirability factors under varying defensiveness conditions. *J. clin. Psychol.*, 1958, **14**, 283–287.

Kelleher, D. The social desirability factor in Edwards' PPS. *J. consult. Psychol.*, 1958, **22**, 100.

Navran, L., & Stauffacher, J. C. Social desirability as a factor in Edwards Personal Preference Schedule performance. *J. consult. Psychol.*, 1954, **18**, 442.

Taylor, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, **23**, 514–517.

(Received April 13, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 205–209

BEHAVIOR PROBLEMS OF MIDDLE CHILDHOOD[1]

DONALD R. PETERSON
University of Illinois

Before the etiology and treatment of children's behavior disorders can be sensibly examined, the disorders themselves must be defined. For the sake of generality and descriptive efficiency, any concepts employed in such definition should be nonarbitrary, unitary, and independent. Factor analytic methods have been employed with salutary effect in the structural definition of adult disorders (e.g., Lorr, Jenkins, & O'Connor, 1955; Rubenstein & Lorr, 1957; Wittenborn, 1951; Wittenborn & Holzberg, 1951), but similar work with the disorders of childhood has only begun (Hewitt & Jenkins, 1946; Himmelweit, 1953). The present study extends and refines this earlier research by factorizing uniformly gathered judgments of problem behavior during the kindergarten and elementary school years, and by examining changes in problem expression during that time.

Subjects and Procedures

In the absence of any accepted theory of structural organization among children's behavior disorders, a sample of problems was chosen by empirical means. The referral problems of 427 representatively chosen cases at a guidance clinic were recorded, and frequencies tabulated for all problems mentioned more than once. Groups of synonymous terms were reduced by eliminating all but the most frequently used expression, and four concepts were discarded because they were conceptually supraordinate to other terms, and hence redundant. Choice among the remaining variables was determined exclusively by the frequency with which they had occurred, and the 58 most common problems were selected for general investigation.

In use, the variables were ordered randomly, assembled in a format requiring ratings of 0 (no problem), 1 (mild problem), or 2 (severe problem), and submitted for completion to 28 teachers of 831 kindergarten and elementary school children in six different schools in Illinois. The choice of school children, rather than clients undergoing treatment for judged disorders, was based on the assumption that most such disorders are extremes of continuous "normal" dimensions, and was determined by the desirability of obtaining uniform data on large numbers of subjects within the age range under consideration. The large sample requirement has been met previously (Hewitt & Jenkins, 1946; Himmelweit, 1953) by recourse to case history information, but the dangers of that expedient seemed greater than those in the present course, and the study was begun in the hope that otherwise unselected school children would present sufficiently numerous, severe problems to warrant sensible analysis and yield meaningful results. Distributions of ratings were generally eccentric, but the effects were reduced by excluding some rarely checked problems (dizziness, soiling, and enuresis, which occurred in less than 3% of the cases, were eliminated), and by pooling judgments of mild and severe problems (ratings of 1 and 2) for all the remaining variables.

For analysis, the sample was divided into four groups: a kindergarten sample ($N = 126$), a first and second grade sample ($N = 237$), a group from the third and fourth grades ($N = 229$), and a fifth and sixth grade sample ($N = 239$). Two teacher ratings were available for each kindergarten child; the number of actual ratings used in the analysis is thus double the N given above for the kindergarten group. Phi coefficients of intercorrelation were computed separately for the four samples. From each correlation matrix, 10 centroid factors were extracted, and from each set of centroid factors 2 were rotated to conform with Kaiser's varimax criterion (Kaiser, 1958).

Human judgment was involved only once in all that analysis—in deciding how many factors to retain for rotation. The decision to keep only two was based on inspection of plots of variance removed by successive centroid factors, and the application of criteria for factor retention developed elsewhere

[1] This study was supported by a grant from the Department of Public Welfare, State of Illinois. I am grateful to Jared T. Lyon, Superintendent of the Hoopeston (Illinois) Public Schools; to Lester J. Grant, Charlotte Meyer, and Hazel Dunivan of the Decatur Public Schools; to Donald W. Dunnan and John Carlino of the Springfield Public Schools; and to Glenn Raymond, Superintendent of Elementary Schools in Watseka, Illinois, for their help in collecting the data. Most of all, I wish to thank the teachers who worked so hard and with such evident care on the ratings themselves.

(Peterson, 1960). Out of personal curiosity, five-factor solutions were also tried for each data set; but these, as expected, were much less stable over age than the two-factor solutions, and only the latter will be reported.

Factor scores were computed for all cases by unweighted summation of pertinent problems checked by the teachers. Interjudge correlations were computed for the kindergarten sample, and further attention directed toward comparing the four age groups. Data for boys and girls were separated in all comparisons, because of well known sex differences in problem expression, and mean factor scores were computed to show trends in the development of behavior problems over the years of middle childhood.

Results

The Factors

All four sets of rotated factor loadings are presented together in Table 1, an arrangement permitted only by the marked similarity between results at the four age levels.[2] Factor 1 is obviously a *conduct problem* dimension, closely resembling the like-named factor isolated by Himmelweit (1953) and "unsocialized aggression" as defined by Hewitt and Jenkins (1946). Factor 2 has been labeled *personality problem* in accordance with Himmelweit's designation and common usage. It is much like the "over-inhibited behavior" dimension which Hewitt and Jenkins found. Actually these terms, "personality problem" and "conduct problem," are grossly inappropriate. Both problems are personality expressions, and both affect conduct. But the central meanings seem clear enough. In one case, impulses are expressed and society suffers; in the other case impulses are evidently inhibited and the child suffers.

The generality of these factors appears to be enormous. Not only do they emerge with striking uniformity over the limited age range and the particular variables and subjects examined here; they have appeared in very much the same form with the recorded problems of treatment cases (Hewitt & Jenkins, 1946; Himmelweit, 1953), and remarkably similar factors have appeared in the questionnaire behavior of delinquent boys (Peterson, Quay, & Cameron, 1959). Considering all studies together, age has varied from early childhood to adolescence; problem status has varied from none, through clinic attendance, to incarceration for delinquency; data sources have varied from case history records, to standard ratings, to questionnaire responses; methods of factor extraction have varied from cluster inspection to centroid analysis; rotational methods have varied from none, through visual shifts to both orthogonal and oblique solutions, to analytic techniques. Through it all, the factors have stayed the same, and their definition at last seems adequate. The time is ripe for study, particularly experimental study, of dynamics, etiology, and treatment.

Factor Scores and Their Reliability

Such investigations, however, cannot proceed until various properties of the measuring devices have been examined. Factor scores were computed by unweighted summation over the first 15 variables for each factor as listed in Table 1. Below that point, many of the variables either have no appreciable loading on either dimension, or approximately equal loadings on both. The former condition holds especially for skin allergy, hay fever, nausea, and stomach-aches, which may often be purely somatic and qualitatively distinct from the other variables examined. The latter condition, roughly equal loadings on both factors, holds for crying, nervousness, and certain school attitudes, variables which are either very general in nature or exhibit some kind of developmental change.

Reliability and interfactor correlation were examined for the kindergarten group only, since only for that group were dual ratings available. Interjudge r's of .77 and .75 were found for Factors 1 and 2, respectively. These figures are exceptionally good for ratings, and are sufficiently high for most research purposes. The correlation between factors was .18, low enough to meet most requirements for independence.

[2] The rating schedule, correlation matrices, and unrotated centroid factor matrices have been deposited with the American Documentation Institute. Order Document No. 6632 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $2.00 for microfilm or $3.75 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

ROTATED FACTOR LOADINGS

Factor	Conduct Problem				Personality Problem			
	K[a]	1–2	3–4	5–6	K[a]	1–2	3–4	5–6
Conduct Problem								
Disobedience	74	77	69	86	03	04	07	11
Disruptiveness	73	67	66	76	−04	19	−03	11
Boisterousness	68	63	67	68	−16	07	−07	−09
Fighting	54	73	61	77	−09	−04	11	07
Attention-seeking	54	67	63	76	−12	10	−07	02
Restlessness	64	58	62	71	04	24	06	20
Negativism	56	64	60	70	12	27	20	15
Impertinence	57	57	53	76	02	−08	00	08
Destructiveness	59	65	51	65	−05	27	19	00
Irritability	53	59	57	69	01	11	04	07
Temper tantrums	54	37	49	64	08	11	22	16
Hyperactivity	51	49	54	49	−06	12	00	03
Profanity	30	42	64	60	−07	11	02	00
Jealousy	23	50	41	56	06	10	12	11
Uncooperativeness	67	67	53	71	09	31	38	21
Distractibility	56	57	61	72	29	42	32	26
Irresponsibility	60	65	49	65	22	18	47	20
Inattentiveness	54	61	36	69	39	30	57	28
Laziness in school	44	59	36	37	29	36	55	31
Shortness of attention span	48	54	31	60	37	34	55	29
Dislike for school	38	40	32	41	06	26	54	13
Nervousness	22	25	46	50	40	44	22	26
Thumb-sucking	29	17	36	05	09	28	−03	15
Skin allergy	−16	02	−20	−05	01	21	05	−20
Personality Problem								
Feelings of inferiority	12	25	13	17	59	56	66	62
Lack of self-confidence	12	26	13	16	60	61	60	58
Social withdrawal	−03	08	04	05	50	64	61	60
Proneness to become flustered	07	28	15	24	54	59	60	58
Self-consciousness	−13	−03	−15	16	55	60	47	63
Shyness	−16	−18	−23	−13	62	57	50	51
Anxiety	01	19	24	10	50	57	55	47
Lethargy	−06	22	01	31	52	47	61	43
Inability to have fun	−15	06	−14	−09	49	48	53	48
Depression	00	20	04	29	47	43	64	42
Reticence	06	20	08	14	45	43	64	41
Hypersensitivity	06	26	18	30	40	53	54	46
Drowsiness	02	09	09	29	39	48	45	41
Aloofness	−16	−03	−04	05	51	32	50	31
Preoccupation	09	12	23	37	47	57	64	41
Lack of interest in environment	24	30	21	51	40	44	67	28
Clumsiness	16	21	36	17	43	54	34	36
Daydreaming	14	26	21	49	53	46	69	47
Tension	21	31	39	39	41	62	27	41
Suggestibility	04	29	31	52	41	42	48	30
Crying	15	14	06	59	27	48	32	19
Preference for younger playmates	28	21	23	14	08	45	37	32
Specific fears	−09	24	−02	−04	24	47	20	20
Stuttering	11	17	08	02	27	35	29	16
Headaches	19	21	00	00	07	46	22	27
Nausea	−10	23	07	−02	01	14	38	37
Truancy from school	27	07	04	22	00	20	39	35
Stomach-aches	10	18	05	−06	−01	30	38	29
Preference for older playmates	−14	05	26	16	01	38	16	01
Masturbation	08	14	26	04	−18	40	04	17
Hay fever or asthma	15	−01	17	−11	05	21	09	03

[a] Kindergarten.

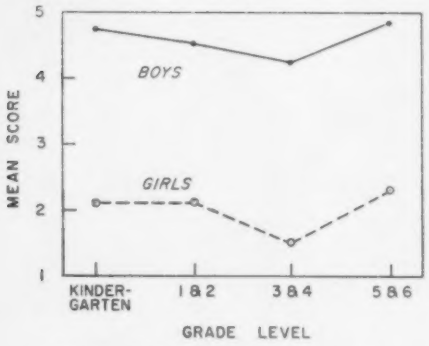

FIG. 1. Mean conduct problem scores.

Developmental Changes

Mean factor scores were computed for boys and girls in all age groups, and the results are shown in Figures 1 and 2. Throughout middle childhood, boys consistently display more severe conduct disturbances than girls, possibly as a function of constitutional differences, but more likely in response to different levels of social expectancy and tolerance for misbehavior. An interesting reversal, however, occurs in the expression of personality problems. Boys evidently start school with more personality problems than girls, but around the seventh or eighth year such problems become more plentiful among girls. Again, social pressures for sex-type conformity seem the likeliest causal agents. Reasons for the apparent upswing in problems at the fifth and sixth grade level are obscure. The increase may arise from the early agitation of adolescence, and the difficulty this can bring about in our society.

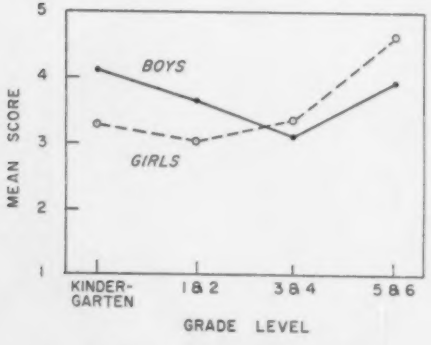

FIG. 2. Mean personality problem scores.

SUMMARY

This study was designed to improve structural definition of children's behavior problems and to examine changes in those problems over the years of middle childhood. Teacher ratings of 58 clinically frequent problems were obtained for 831 kindergarten and elementary school children, and four separate factor analyses were conducted, one for the kindergarten subjects and one each for children in grades 1–2, 3–4, and 5–6. Two factors emerged with remarkable invariance in all four analyses. The first implied a tendency to express impulses against society, and was labelled "conduct problem." The second contained a variety of elements suggesting low self-esteem, social withdrawal, and dysphoric mood. It was called "personality problem." Both factors have now appeared in a number of studies despite wide differences in subjects, variables, and anlaytic procedures.

Comparisons over age showed that boys displayed more severe conduct problems than girls at all age levels examined. Kindergarten and primary school boys also showed more severe personality problems than girls, but at the two highest age levels this trend was reversed, and girls displayed more personality problems than boys.

The definition of both dimensions seems adequate. Reliable, independent measures of the factors can be obtained, and the way toward investigation of dynamics, etiology, and treatment now seems clear.

REFERENCES

HEWITT, L. E., & JENKINS, R. L. *Fundamental patterns of maladjustment: The dynamics of their origin*. Springfield, Ill.: Green, 1946.

HIMMELWEIT, HILDE T. A factorial study of "children's behavior problems." Cited in H. J. Eysenck, *The structure of human personality*. London: Metheun, 1953.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, 23, 187–200.

LORR, M., JENKINS, R. L., & O'CONNOR, J. P. Factors descriptive of psychopathology and behavior of hospitalized psychotics. *J. abnorm. soc. Psychol.*, 1955, **50**, 78–86.

PETERSON, D. R. The age generality of personality factors derived from ratings. *Educ. psychol. Measmt.*, 1960, **20**, 461–474.

PETERSON, D. R., QUAY, H. C., & CAMERON, G. R. Personality and background factors in juvenile delinquency as inferred from questionnaire responses. *J. consult. Psychol.*, 1959, **23**, 395–399.

RUBENSTEIN, E. A., & LORR, M. Patient types in outpatient psychotherapy. *J. clin. Psychol.*, 1957, **13**, 356–361.

WITTENBORN, J. R. Symptom patterns in a group of mental hospital patients. *J. consult. Psychol.*, 1951, **15**, 290–302.

WITTENBORN, J. R., & HOLZBERG, J. D. The generality of psychiatric syndromes. *J. consult. Psychol.*, 1951, **15**, 372–380.

(Received April 18, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 210–215

ATTRIBUTION OF TRAITS AND EMOTIONAL HEALTH AS FACTORS ASSOCIATED WITH THE PREDICTION OF PERSONALITY CHARACTERISTICS OF OTHERS[1]

MARVIN SPANNER[2]
University of California, Berkeley

Many unsystematized explanations have been offered concerning the manner in which judges predict the personality characteristics of others. The present paper will focus on two theories currently popular. The first, based on an interpersonal theory of personality, has been most extensively represented by Harry Stack Sullivan. Sullivan's (1947) ideas on the subject are reflected in his famous dictum:

"It is not that as ye shall judge so shall ye be judged, but as you judge yourself, so shall you judge others; strange but true so far as I know, and with no exception."

The above quotation suggests that interpersonal prediction is based primarily on an attributive mechanism in which the state of the individual's self-concept determines the quality of the interpersonal appraisal. One can deduce from his statement the following hypothesis:

I. There is a positive relationship between accuracy in predicting the personality characteristics of others and similarity of personality characteristics of the judge and the individual being judged. This would logically appear to follow because an understanding of others is based, according to Sullivan, on one's self. Thus, if one tends to perceive others as similar to one's description of one's self, then those actually similar will be perceived accurately. Lingren and Robinson (1953) make a similar point when they state, in criticizing Dymond's study (1950), that "Conventional people get good scores on empathy tests because most of their partners (or referrants) in the test are also conventional."

The second group of explanations of predictive ability involves the mental health of the individual. Cline (1953) in an exhaustive review of the literature suggests that, on the whole, a good judge of others is emotionally sound, has good interpersonal relations, is happier, more popular and flexible. In addition, most clinicians assume that an emotionally healthy individual is a better judge of personality because he has fewer problems which interfere with his understanding of others. In the present instance, two measures have been utilized to evaluate the emotional health of the subjects. The first is the correlation between the judge's evaluation of himself and his evaluation of the sort of individual he would ideally like to be. The second is a composite rating by experienced observers of the "soundness" of the judge.

[1] This paper is based in part on a thesis submitted in partial fulfillment of the requirements for the PhD in the Department of Psychology of the University of California, Berkeley. The writer wishes to express his appreciation and indebtedness to Harrison G. Gough, thesis advisor; Donald W. MacKinnon, director of the Institute of Personality Assessment and Research; and Victor B. Cline.

This research was supported in part by the United States Air Force under Contract No. AF 18 (600)-8, monitored by Technical Director, Detachment No. 7 (Officer Education Research Laboratory), Air Force Personnel and Training Research Center, Maxwell Air Force Base, Alabama. Permission is granted for reproduction, translation, publication, use, and disposal, in whole and in part, by or for the United States Government. Personal views or opinions expressed or implied in this publication are not to be construed as necessarily carrying the official sanction of the Department of the Air Force or of the Air Research and Development Command.

[2] Now at Neuropsychiatric Institute, University of California, Los Angeles.

We have, therefore, the following two hypotheses relating the "mental health" of the judge and his ability to predict personality characteristics of others:

II. There is a positive relationship between the self–ideal-self correlation of the judge (called self-acceptance and used as a measure of mental health) and his ability to predict personality characteristics of others.

III. There is a positive relationship between the rated "soundness" of a judge (used as a measure of mental health) and his ability to predict personality characteristics of others.

Method

Since both the tests of judging ability and the prediction instruments were devised and fully reported by Cline (1955), the following is an abbreviated description of these procedures.

Tests of Judging Ability

The basic procedure used was a sound movie of four fictitious employment interviews. The interviewees were male undergraduates. Three of them were 19 years old and single. The fourth was 33 years old, married, childless, and a veteran. Each observer or judge was asked, after viewing each one of the four interviews, to make a series of predictions about the interviewee (hereafter called social object) on three specially devised instruments, only one of which was used in the present analysis.

The four filmed interviews, conducted by a trained actor, were highly structured and fairly constant. These were chosen from a group of nine films in order to obtain the most diverse personalities among the social objects. Cline has stated that the interviews are divided into three phases: a standard job interview situation, a stress situation in which the interviewer is highly critical of the interviewee, and a relaxed after-interview abreaction session in which the interviewee's reactions to the interview are discussed. The intent was to get a dynamic record of a series of subjects on a sound film, rich and varied enough in the use of verbal and visual cues so that a judge would have an adequate sample of each social object's social technique.

Prediction Task

The sole prediction task used in the present study to assess the judging ability of persons who observed the films was the Personality Word Card (PWC). Here the judge was asked to predict the responses of each of the four social objects on a 100-word adjective check-list. This list was derived from a 300-word Adjective Check-List compiled by Gough.[3]

[3] Gough, H. G. Predicting success in graduate training. (Progress report) Unpublished manuscript, University of California Institute of Personality Assessment and Research, 1952.

The corrected split-half reliability for the PWC as a test of judging ability for a group of 100 college undergraduates was .83 ± .06.

The PWC was also used by both the judges and social objects to describe themselves. Comparisons were, therefore, readily made among the judges' prediction of the social object, the social objects' self-reports, and the judges' self-reports.

Two other prediction tasks were given to the judges, although they were not used in the present analysis. One of them, the Behavioral Postdiction Test, in which the judge was required to predict the social object's behavior in real life, was discarded because it was discovered, when a simple analysis of variance was carried out, that the source of variation attributable to the judges was not significant. A third instrument, a multiple-choice sentence completion test, was also discarded as a test of judging ability, because Cline had found that the corrected split-half reliability for a group of 100 college undergraduates was only .36 ± .06.

Judges

One hundred Air Force Captains were tested at the Institute of Personality Assessment and Research, as part of a large scale study of officer effectiveness. They were tested in groups of 10 in a living-in situation similar to that employed by the OSS during World War II (Office of Strategic Services, 1948). Their average age was 33.6 with a range of 27 to 49. Ninety-five percent of the officers were married, with an average of two children. Fifty-six had attended college, seven had done some graduate work, and only three had not graduated from high school.

Measures

Accuracy. This variable is defined as the tetrachoric correlation between the social objects' self-reports and the judges' prediction of them on the PWC. Since there are four social objects, each one of the 100 judges has four accuracy scores.

Similarity. This measure is defined as the tetrachoric correlation between the social objects' self-reports and the judges' self-reports on the PWC. Here too, each one of the 100 judges has a similarity score for each of the four social objects.

Self-Acceptance (self–ideal-self correlation). Each judge completed the PWC twice, once describing himself as he was and then a second time describing himself as he would ideally like to be. The measure of association used was the phi coefficient. It is assumed that the greater the degree of association the more self-accepting and therefore the "healthier" the individual.

Soundness. This variable was drawn from a pool of 30 traits which were used by 10 staff members in rating the judges (officers). The following procedure was used. After the first four groups of 10 officers had been run, 10 staff members of the Institute of Personality Assessment and Research rated each of the 40 officers in the 30 traits. A normal distribution was required of the raters, using

TABLE 1

CORRELATION COEFFICIENTS (PEARSONIAN) BETWEEN ACCURACY AND SIMILARITY, SELF-ACCEPTANCE, AND SOUNDNESS, FOR EACH OF THE FOUR SOCIAL OBJECTS

Measure	Accuracy in Predicting the Four Social Objects			
	I	II	III	IV
Similarity	.01	.28*	.00	.00
Self-Acceptance	−.06	.15	.06	−.03
Soundness	−.10	−.13	.13	−.09

* $p < .01$.

a five-point scale. The same procedure was followed in rating the remaining 60 judges, after they had all been assessed. A composite trait rating was derived from each assessee by combining the individual ratings of the 10 staff members. Each item of the trait pool was then assigned a scale value for each assessee.

Soundness was defined as:

> Maturity in personal relations; self-insight and self-acceptance, as well as acceptance and understanding of others. Absence of serious emotional problems. Stability of mood and manner. Good balance of social conformity and spontaneity.

All of the above measures, once obtained, are simply treated as scores, for purposes of statistical analysis.

RESULTS

The first hypothesis—that accuracy of prediction of personality traits of others is dependent on the degree of similarity of judge and social object—was tested by correlating similarity and accuracy for each of the social objects. As can be seen in Table 1, the correlation between similarity and accuracy for Social Object I is .01 and for Social Objects III and IV is .00. The correlation for Social Object II, however, is .28, significant at the .01 level of confidence. Since this is the only one of the four social objects where significance was achieved, the first hypothesis, that there is a relationship between accuracy of prediction and similarity of judge and social object, was not substantiated.

The following results are concerned with the relationship of mental health and the ability to predict personality characteristics of others. As can be seen in Table 1 there is no significant relationship between the ability to predict personality characteristics of others and self-acceptance (self–ideal-self correlation) or soundness of the judge, and, therefore, both Hypotheses II and III have not been substantiated.

Although it has been shown that similarity of personality characteristics of judge and social object is not related to accuracy of prediction on the part of the judge, other factors might prevent a simple expression of this relationship. For example, a judge who is similar to a social object may predict accurately only if he is self-accepting (i.e., his self–ideal-self correlation is high). With this in mind, a triple classification analysis of variance for judging ability was carried out. The judges were divided into a high and low similarity group as well as a high and low self-acceptance group (self–ideal-self correlation), for each of the four social objects. Judges above the median were considered high and those below the median were considered low on both similarity and self-acceptance.

Because of the varying number of observations in the cells of the triple classification, it was necessary to pool the residual and triple interaction sums of squares. This is a conservative estimate of the error variance; if the triple interaction is significant, pooling it with the error variance would only produce an overestimate of this term and tend to produce smaller F ratios.

TABLE 2

ANALYSIS OF VARIANCE OF JUDGING ABILITY, IN TERMS OF SIMILARITY AND SELF-ACCEPTANCE, FOR THE FOUR SOCIAL OBJECTS

Source of Variation	*SS*	*df*	*MS*	*F*
Similarity (Sim)	656	1	656	3.14
Self-Acceptance (SA)	167	1	167	.80
Social Objects (SO)	92,219	3	30,740	147.08*
Sim × SO	957	3	319	1.53
SA × SO	761	3	254	1.22
Sim × SA	3,693	1	3,693	17.67*
Sim × SA × SO and Residual	80,836	387	209	
Total	179,289	399		

* $p < .01$.

TABLE 3

ACCURACY AND "FAVORABLE" ADJECTIVES FOR EACH OF THE SOCIAL OBJECTS

Social Object	Mean Accuracy	Percentage of Gough's "Favorable" Adjectives Used by Social Objects in Their Self-description
I	46.24	42
II	23.96	30
III	33.51	31
IV	64.51	48

The analysis of variance shown in Table 2 reveals two significant mean squares: social objects, and the interaction between similarity and self-acceptance. In attempting to understand the basis for the large mean square for social objects, it was noted that the self-descriptions of the social objects tended to be quite favorably toned. A measure of this favorable tone was derived from Gough's list of "favorable" adjectives. Gough (see Footnote 3) was able to derive the list by asking a group of 30 college students to rate the entire list of adjectives for its favorability. The highest rated 25% were selected for the favorability key. One question which immediately comes to mind is whether there is a relationship between accuracy of prediction, on the part of the judge, and favorableness of self-perception, on the part of the social object. If one compares the list of favorable adjectives with the mean accuracy score for each of the social objects, in Table 3, one notes a perfect correlation (rho) between the two variables. Thus, the accuracy of prediction appears to depend, in part, on favorableness of self-perception of the social object.

The second significant mean square noted in Table 2 is the interaction between similarity and self-acceptance. To further clarify the relationship, the accuracy scores of judges conforming to the four possible combinations of similarity and self-acceptance, under the present conditions, are shown in Table 4. Two groups can be considered relatively accurate: judges similar to a social object and self-accepting, and judges dissimilar to a social object and not self-accepting. The other two groups can be considered relatively inaccurate: judges similar to a social object and not self-accepting, and judges dissimilar to a social object and self-accepting.

DISCUSSION AND CONCLUSIONS

What are the implications of Sullivan's statement that one judges others only in terms of one's self? It obviously creates a huge solipsism in which no distinction is drawn between one's self and others in the judgment process. The lack of relationship exhibited between similarity of judge and social object and accuracy of prediction (Hypothesis I), in the present experiment, suggests that the prediction of personality characteristics of others is not a reflection of the judge, a simple attribution to another of the traits of the judge.

In addition, no relationship was found, in the present study, between the mental health of an individual—using (*a*) a self-rating criterion, self-acceptance (self–ideal-self correlation); and (*b*) an external criterion, judged soundness of an individual by expert raters—and his ability to predict personality characteristics of others (Hypotheses II and III).

When, however, both similarity and self-acceptance (self–ideal-self correlation) interact, there is a significant effect on accuracy. This result can be understood if we make some assumptions concerning the psychological meaning of each of these variables. Let

TABLE 4

MEAN ACCURACY OF THE INTERACTION ON HIGH AND LOW SCORES ON BOTH SIMILARITY AND SELF-ACCEPTANCE

Group	*N*	Mean	*SD*
High Sim and High SA	120	46.18	20.07
High Sim and Low SA	80	39.08	22.40
Low Sim and High SA	80	37.49	21.23
Low Sim and Low SA	120	42.97	21.93

Note.—High = Judges above median on Sim or SA; Low = judges below median.

us assume that a judge who is relatively self-accepting will tend to judge others as similar to himself. The unconscious basis for his predictions might be as follows: "I am a nice person and so are other people." A judge who is relatively low on self-acceptance, however, will tend to judge others as dissimilar to himself. Here, the unconscious basis for his predictions might be: "I am not a nice person, but others are." The mechanisms described here are similar to those characterized by Cameron and Magaret (1951) as, respectively, assimilative and disowning projection.

As was described above, there are objective measures of the similarities of judge and social object. If we focus our attention on those judges who are high on self-acceptance, we can hypothesize that they would be accurate in the prediction of personality characteristics of those social objects similar to themselves, and inaccurate in the prediction of personality characteristics of those social objects dissimilar to themselves. This would logically follow since, as postulated above, a self-accepting judge would tend to perceive others as similar to himself. Those actually similar to him would, therefore, be judged accurately, while those dissimilar to himself would be judged inaccurately.

In a similar fashion, we would predict that judges with low self-acceptance scores would accurately perceive those social objects who are dissimilar to themselves, while inaccurately perceiving those social objects similar to themselves. This, too, would logically follow since, as postulated above, judges with low self-acceptance scores would tend to perceive others as dissimilar to themselves. Those social objects actually similar to these judges would, therefore, be perceived inaccurately, while those dissimilar to them would be perceived accurately.

The post hoc explanation invoked here conforms to the results in Table 4. While the results presented suggest that judging the personalities of others involves more than an attribution of one's traits to another, it also suggests that judges engage in a mechanical type of prediction based on their self-concept. If the judges are self-accepting, they judge others in a similar manner. If the judges are not self-accepting, they judge others in a contrary manner. Accuracy, therefore, depends not only on whether an individual is self-accepting or not, but also on whether the judges are similar or dissimilar to the social object. Bronfenbrenner (1958) has made a similar point. He utilized the variable "Favorability toward others" in his analysis, however, rather than self-acceptance used in the present study.

Looked at from this point of view, the ability to predict personality characteristics of others should not be perceived as a trait residing within the judge, but, rather, as based on the personality patterns of both judge and social object. This point of view is further substantiated by an analysis of variance of judging ability carried out by the author (Spanner, 1955), in which the interaction between judge and social object approached significance at the .05 level.

This approach suggests, therefore, that the question to be asked will no longer exclusively be "Is he a good judge of personality?" but also "Which social objects can he judge well?" Another facet of the same problem would be an attempt to break down the evaluation of personality into various areas and to assess the ability of judges in each area. This presupposes a representative design of the sort suggested by Brunswik (1947), in which there would be not only representative sampling of subjects (judges), but also of objects (social objects). In the present instance a preliminary attempt was made to analyze the different skills involved in judging social objects with favorable as compared to unfavorable self-perceptions. The attempt was abandoned, however, because of the small number of social objects (four).

Possibly if some of the aforementioned approaches are handled successfully, eventually, we will then be in a position to answer the following question: "How does the ability of a judge to predict the personality characteristics of others relate to his own personality dynamics?" This, of course, is related to the entire area of ego psychology and the differential utilization of ego defenses by an individual in judging the personality characteristics of others.

SUMMARY

One hundred military officers were presented with sound movies of stress interviews with four interviewees (social objects) and asked to predict the responses of the social objects on an adjective check-list, as well as to describe themselves on the same instrument. From these basic data measures of accuracy of prediction and similarity of judge and social object were obtained. Measures of the self-acceptance (self–ideal-self correlation) and "soundness" of the judges were also obtained.

Contrary to expectations, accuracy was found to be unrelated to similarity of judge and social object, to self-acceptance, and to the rated soundness of the judge. Accuracy scores were related, however, to a combination of similarity and self-acceptance variables. The concepts of disowning and assimilative projection were utilized as a post hoc explanation of accuracy.

REFERENCES

BRONFENBRENNER, U. The measurement of skill in social perception. In D. C. McClelland, A. L. Baldwin, U. Bronfenbrenner, & F. L. Shadtbeck (Eds.), *Talent and society*. Princeton: Van Nostrand, 1958.

BRUNSWIK, E. Systematic and representative design of psychological experiments. *U. Calif. Syllabus Ser.*, 1947, No. 304.

CAMERON, N., & MAGARET, ANN. *Behavior pathology*. Boston: Houghton Mifflin, 1951.

CLINE, V. B. The assessment of good and poor judges of personality using a stress interview and sound film technique. Unpublished doctoral dissertation, University of California, Berkeley, 1953.

CLINE, V. B. Ability to judge personality assessed with a stress interview and sound-film technique. *J. abnorm. soc. Psychol.*, 1955, **50**, 183–187.

DYMOND, ROSALIND F. Personality and empathy. *J. consult. Psychol.*, 1950, **14**, 343–350.

LINGREN, H. C., & ROBINSON, JACQUELINE. An evaluation of Dymond's test of insight and empathy. *J. consult. Psychol.*, 1953, **17**, 172–176.

OFFICE OF STRATEGIC SERVICES, Assessment Staff. *Assessment of men*. New York: Rinehart, 1948.

SPANNER, M. Similarity, identification, and distortion as factors in the prediction of personality characteristics. Unpublished doctoral dissertation, University of California, 1955.

SULLIVAN, H. S. *Conceptions of modern psychiatry*. Washington, D. C.: William Alanson White Psychiatric Foundation, 1947.

(Received April 18, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 216–220

HOSTILITY EXPRESSION AMONG DELINQUENTS OF MINORITY AND MAJORITY GROUPS

DON L. SWICKARD AND BERNARD SPILKA
University of Denver

The frustration-aggression hypothesis has frequently provided a theoretical framework for explaining the phenomena of crime, delinquency (Dollard, Doob, Miller, Mowrer, Sears, Ford, Hovland, & Sollenberger, 1939), and prejudice (Allport, 1954; Dollard et al., 1939). Two prevalent and pervasive sources of frustration which are seen as motivating anti-social behavior are low socioeconomic status and minority group membership (Dollard et al., 1939). Evidence to support this position may be adduced from the work of Glueck and Glueck (1950) and Sutherland and Cressy (1955), who show clearly the relationship between poverty and delinquency. On the other hand, Hammer (1953), McCary (1950), and Mussen (1953) have demonstrated that minority group members reveal more manifest aggression than do members of majority groups. The factors of poverty and discrimination have been confounded in these studies, thus failing to clarify the role of each. It may be hypothesized, however, that minority group membership and low socioeconomic status combine to effect a greater manifestation of hostility than would simply result from the frustrations of poverty alone. Delinquents who are members of a minority might, therefore, be expected to reveal more evidences of hostility than delinquents who are majority group members. The present investigation was designed to assess this position, utilizing Spanish-American and non-Spanish, white delinquents. That the Spanish-American group is a negatively valued minority has been pointed out by Jones (1948) and Saunders (1954) in their writings concerning the Spanish- or Mexican-American, and his status among Anglo-Americans. Specifically, the hypotheses tested were:

1. A group of Spanish-American delinquents of low socioeconomic status will show significantly greater evidences of aggression, in both amount and kind, than will be demonstrated by a similar group of non-Spanish, white delinquents.

2. Because of tendencies on the part of delinquents to give a "good impression" on personality measures (Lindzey & Goldwyn, 1954; Vane, 1954), it will be necessary to correct hostility measures for social desirability.

3. Social desirability will relate negatively to manifest and extrapunitive measures of hostility and positively to intropunitive and impunitive measures of hostility.

METHOD

Subjects. The original group studied consisted of 81 male and female delinquents on probation at the Denver Juvenile Court. The subjects were selected in such a way as to eliminate all who were the result of mixed group backgrounds, who had been diagnosed as neurotic or psychotic, were other than lower class, or who had been institutionalized in the last 2 years. On the basis of obtaining scores of 10 or more on the MMPI Lie scale, 7 subjects were eliminated. The final group thus consisted of 25 Spanish-American males, 12 Spanish-American females, 25 non-Spanish white males, and 12 non-Spanish white females. All subjects were between 14 and 17 years of age.

Tests. The tests administered were the Rosenzweig Picture Frustration Study (Rosenzweig, Fleming, & Clarke, 1947), the Siegel Manifest Hostility scale (Siegel, 1956), the 39-item Social Desirability scale extracted from the MMPI by Edwards (1957), and the MMPI Lie scale (Hathaway & McKinley, 1951).

Procedure. All subjects were tested in groups of from 5 to 15, in the courtroom of the Denver Juvenile Court. All subjects were assured that the testing had nothing to do with their probation and

that the results would not be made available to their probation counselors. All subjects were cooperative and answered all questions without difficulty.

Results and Discussion

Table 1 contains the means and standard deviations for the four groups on the various tests administered. Pearson product-moment correlations were computed between the Social Desirability scale and the other measures, as seen in Table 2. Hypotheses 2 and 3 obtained support as the Social Desirability scale correlated −.655 with the Siegel Manifest Hostility scale, −.372 with the extrapunitive scores, .222 with the intrapunitive scores, and .262 with the impunitive scores. The first two coefficients are significant at the .01 level of confidence, and the last two at the .05 level, with 72 degrees of freedom.

One-way simple classification analyses of variance and covariance were computed between the four groups: Spanish-American males and females, and non-Spanish white males and females. This analysis was more conservative than a group × sexes design would have yielded. The latter was computed in the analysis of variance for a 2 × 2 disproportionate subclass numbers design as recommended by Snedecor (1956), in order to check for further significance than might have been obtained via the single variable design. No additional significance was obtained for the group, sex, or interaction terms. Because of the problem of interpreting an analysis of covariance for the 2 × 2 design with unequal and disproportionate subclass numbers, it was decided to employ the more conservative analysis reported here.

The analyses of variance across groups were significant at the .05 level for social

TABLE 1

Means and Standard Deviations of Age, Social Desirability, and Manifest Hostility and of Extrapunitive, Intropunitive, and Impunitive Scores for All Groups

Variable[a]	Group[b] SA-M (N = 25)	SA-F (N = 12)	NSW-M (N = 25)	NSW-F (N = 12)	Total (N = 74)
Age					
Mean	15.0	14.7	15.0	14.5	14.8
SD	.72	.75	.80	.76	.76
SD					
Mean	26.2	20.2	23.6	24.4	24.0
SD	5.10	4.88	6.18	3.59	5.62
MH					
Mean	22.2	23.4	19.6	19.3	21.1
SD	6.62	6.49	8.01	9.14	7.71
E					
Mean	8.5	9.2	8.2	9.0	8.6
SD	2.99	3.96	3.58	3.20	3.42
I					
Mean	7.1	6.4	6.5	6.6	6.7
SD	2.33	2.17	2.44	2.59	2.40
M					
Mean	7.1	6.5	7.7	7.0	7.2
SD	2.23	3.34	2.77	1.58	2.58

[a] *SD* refers to Social Desirability scale; *MH*, Manifest Hostility scale; E, Extrapunitive Scores; I, Intropunitive Scores; M, Impunitive Scores.

[b] SA-M refers to Spanish-American males; SA-F, Spanish-American females; NSW-M, non-Spanish, white males; NSW-F, non-Spanish, white females.

TABLE 2

CORRELATIONS BETWEEN THE SOCIAL DESIRABILITY SCALE AND THE MANIFEST HOSTILITY SCALE, AND THE EXTRAPUNITIVE, INTRAPUNITIVE, AND IMPUNITIVE SCORES ON THE ROSENZWEIG PICTURE–FRUSTRATION STUDY

Variable	Group				
	SA-M ($N = 25$)	SA-F ($N = 12$)	NSW-M ($N = 25$)	NSW-F ($N = 12$)	Total[a] ($N = 74$)
SD and *MH*	−.648**	−.627*	−.706**	−.666*	−.655**
SD and E	−.110	−.581	−.466	−5.05	−.372**
SD and I	−.136	.421	.288	.581*	.222*
SD and M	.104	.588*	.772**	.308	.262*

[a] The *r*'s for the total group are not based on an averaging of the subgroup coefficients but on direct computation from the basic data. All subgroup coefficients in the first three rows are homogenous. Those in Row 4 are heterogenous and the total *r* of .262 is at best a conservative estimate of this relationship.
* Significant at .05 level.
** Significant at .01 level.

desirability (see Table 3), and not significant for the Manifest Hostility scale (see Table 3). No significance was obtained for the analyses of extrapunitiveness, intropunitiveness, and impunitiveness (see Table 3). Because of the significant differences among the groups in social desirability and the correlations noted above, analyses of covariance were computed among the groups on the hostility scores, while adjusting the means for social desirability. Significance was attained for the Manifest Hostility scores, but not for any of the other measures (see Table 3). The adjusted means computed for the Manifest Hostility scale are: Spanish-American males, 24.3; Spanish-American females, 19.8; non-Spanish white males, 19.3; and non-Spanish white females, 19.7.

It is apparent that Hypothesis I is partially supported since the Spanish-American male group obviously manifested more hostility on the Manifest Hostility scale than did any of the other groups.

TABLE 3

F RATIOS BETWEEN GROUPS FOR ORIGINAL MEANS AND MEANS ADJUSTED FOR SOCIAL DESIRABILITY

Variable	Original Means ($df = 3.70$)	Adjusted Means ($df = 3.69$)
SD	3.327*	–
MH	1.011	3.240*
E	<1	<1
I	<1	<1
M	<1	<1

* Significant at .05 level.

It is not clear why the Spanish-American females did not reveal more evidences of hostility than did either of the non-Spanish white groups. It is noteworthy that the significance obtained among the groups on the Social Desirability scale is probably, in part, due to the fact that the Spanish-American females received the lowest mean score on this variable.

Jones (1948) in a study of the ethnic patterns of the Mexican family in the United States indicates that the home training given girls in this group is radically different from that given boys. Among other things manifestations of hostility are strongly disapproved. The boys of this group might thus be more similar to the non-Spanish white group in background, and it is possible that the Social Desirability scale is thus less appropriate for the Spanish-American female than for the male. This somewhat different conception of social desirability plus pressures against evidencing overt aggression would account partially for the low Social Desirability scale scores obtained by the girls, and failure of the adjusted Manifest Hostility scores to increase along with those of the Spanish-American males.

The failure of the Rosenzweig Picture-Frustration Study to yield significance has been noted before (Lindzey & Goldwyn, 1954;

Vane, 1954), and it may be that the P-F study is actually not measuring manifest hostility but rather aggression on a different level of the personality than that which might be significant for the present study. The P-F study does, however, seem sensitive to the variable of social desirability, as is the Siegel Manifest Hostility scale, and it is suggested that future research involving these two instruments take this into consideration, as was done here.

Another meaningful consideration may be based on the possibility that the lower-class culture simply supports the expression of hostility without it necessarily being related to frustration. The assumed frustration-aggression relationship may therefore be inappropriate. This would account for the differences observed on the Manifest Hostility scale and the failure of the P-F study to reveal the hypothesized tendencies since the latter measure assumes the significance of frustration.

In brief, the results of the present investigation indirectly lend support to the frustration-aggression hypotheses by suggesting that the combination of frustration from low socioeconomic status and minority group membership may increase the expression of aggression over that which is observed when only low socioeconomic conditions are present.

Summary and Conclusions

The present study was designed to investigate the relation of hostility to a combination of low socioeconomic status and minority group membership. The frustration-aggression hypothesis was employed as a theoretical referent, and it was hypothesized that the increased frustration of minority group membership in addition to low socioeconomic status, would produce more manifestations of hostility than would be observed in majority group members of similar class levels. It was further hypothesized that tendencies to give a "good impression" would relate negatively to manifest hostility and extrapunitive tendencies while relating positively to intropunitiveness and impunitive expressions. On the basis of the latter considerations, it was felt that obtained hostility scores would have to be adjusted to minimize such biases.

Eighty-one Spanish-American and non-Spanish white delinquents on probation served as subjects. All subjects were administered the Siegel Manifest Hostility scale, the Social Desirability scale derived by Edwards, the Rosenzweig Picture-Frustration Study, and the Lie scale from the MMPI. Seven subjects who obtained scores of 10 or more on the Lie scale were eliminated from the study. Significant negative correlations were found between the Social Desirability scale and the Siegel Manifest Hostility scale, and the extrapunitive scores from the Rosenzweig Picture-Frustration Study. Significant positive correlations were obtained between the Social Desirability scale and the measures of intropunitiveness and impunitiveness. Once the hostility means were adjusted to remove the effects of social desirability, significance was obtained between the groups on the Manifest Hostility scale. The Spanish-American male group was shown to manifest significantly greater hostility on this measure than any other group, thus partially supporting the main hypothesis.

References

Allport, G. W. *The nature of prejudice.* Cambridge: Addison-Wesley, 1954.

Dollard, J., Doob, L. W., Miller, N. E., Mowrer, O. H., Sears, R. R., Ford, G. S., Hovland, C. I., & Sollenberger, R. T. *Frustration and aggression.* New Haven: Yale Univer. Press, 1939.

Edwards, A. L. *The social desirability variable in personality assessment.* New York: Dryden, 1957.

Glueck, S., & Glueck, Eleanor. *Unraveling juvenile delinquency.* New York: Commonwealth Fund, 1950.

Hammer, E. F. Frustration-aggression hypothesis extended to socio-racial areas: Comparison of Negro and white children's H-T-P's. *Psychiat. Quart.*, 1953, **27**, 597–607.

Hathaway, S. R., & McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory.* (Rev. ed.) New York: Psychological Corp., 1951.

Jones, R. C. Ethnic family patterns: The Mexican family in the United States. *Amer. J. Sociol.*, 1948, **53**, 450–453.

Lindzey, G., & Goldwyn, R. M. Validity of the Rosenzweig Picture-Frustration Study. *J. Pers.*, 1954, **22**, 519–547.

McCary, J. L. Ethnic and cultural reactions to frustration. *J. Pers.*, 1950, **18**, 321–326.

Mussen, P. H. Differences between the TAT responses of Negro and white boys. *J. consult. Psychol.*, 1953, **17**, 373–376.

ROSENZWEIG, S., FLEMING, E. E., & CLARKE, HELEN JANE. Revised scoring manual for the Rosenzweig Picture-Frustration Study. *J. Psychol.*, 1947, **24**, 165–208.

SAUNDERS, L. *Cultural differences and medical care.* New York: Russell Sage Foundation, 1954.

SIEGEL, S. M. The relationship of hostility to authoritarianism. *J. abnorm. soc. Psychol.*, 1956, **52**, 368–372.

SNEDECOR, G. W. *Statistical methods.* Ames, Iowa: Iowa State College Press, 1956.

SUTHERLAND, E. H., & CRESSY, D. R. *Principles of criminology.* Philadelphia: Lippincott, 1955.

VANE, JANET R. Implication of the performance of delinquent girls on the Rosenzweig Picture-Frustration Study. *J. consult. Psychol.*, 1954, **18**, 414.

(Received April 18, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 221–225

THE USE OF OPPOSITE SEX SCALES AS A MEASURE OF PSYCHOSEXUAL DEVIANCY[1]

B. G. ROSENBERG, B. SUTTON-SMITH
Bowling Green State University
AND E. MORGAN
Monroe County Schools

Most self-report tests of personality obtain an estimate of male and female sex role identification through the use of like-sex scales. Sex role identification or the lack of it is typically evaluated by the subject's responses to a scale of items constructed and validated on his own sex (Hathaway & McKinley, 1943; Strong, 1943). Thus, a person obtaining an average or high score on his own sex scale is said to be identified with his own sex, and a person obtaining a low score on his own sex scale is said to lack identification with his own sex, or to be more like the opposite sex in his responses. In an earlier paper the authors described the derivation of a scale for boys and girls between the ages of 8 and 12 years in which each subject received a score on an opposite-sex scale as well as on a like-sex scale (Rosenberg & Sutton-Smith, 1959). The existence of these scales for measuring masculinity and femininity has made it possible to investigate the relative effectiveness of both the like-sex scale and the opposite-sex scale for discriminating sex role identification.

It is generally assumed that individuals who are faulty or aberrant in their sex role identification will be more emotionally disturbed than those who are not (Henry, 1948; Terman & Miles, 1936). Therefore, they should tend to have less satisfactory scores on measures of emotional stability. This paper examines the relationship between various types of scores on a masculinity and a femininity scale and several independent measures of emotional stability. The independent measures of emotional stability make it possible to explore not only the existence of sex role confusion, but also certain of its qualitative aspects.

[1] This study was facilitated by a grant from the Scholarly Advancement Committee, Bowling Green State University.

METHOD

A group of 337 children in the fourth, fifth, and sixth grades in two elementary schools in Northwest Ohio and Southeast Michigan comprised the sample.[2] Several instruments were administered to the entire group: (*a*) a check list of games and play activities which is described elsewhere (Rosenberg & Sutton-Smith, 1959) and which effectively yields measures of masculinity and femininity, (*b*) an empirically derived scale which has been found to measure extremes of impulsive behavior (Sutton-Smith & Rosenberg, 1959), (*c*) the Children's Manifest Anxiety scale which affords a measure of anxiety (Castaneda, McCandless, & Palermo, 1956), and finally (*d*) the Brown Personality Inventory (Brown, 1934). The latter scale, derived in 1934, has been found to discriminate neuroticism in children, and to be highly correlated with the *CMA* scale (Rosenberg, Sutton-Smith, & Morgan, in press). Its addition provides subscores which reveal the source of major concern for neurotic children (home, school, physical, insecurity, and irritability). The "home" subscale includes feelings of parental rejection, parental severity, and sibling jealousy. The "school" subscale identifies feelings of inadequacy in the classroom. The "physical" subscale deals with disturbances in physical well-being which appear to be psychogenic in origin. The "insecurity" subscale contains items reflecting anxiety, interpersonal inadequacy, and generalized neurotic concerns. The "irritability" subscale examines emotional reactivity, low tension-binding qualities, and severe feelings of rejection (Brown, 1934).

An analysis was conducted of the scores on impulsiveness, anxiety, and neuroticism of boys and girls in the upper and lower quartiles on the masculinity and femininity scales.

[2] The authors wish to express their indebtedness to H. Lehtomaa, R. Knestrict, and the teachers of Dundee, Michigan, schools for their assistance in the collection of the data.

TABLE 1

NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE MASCULINITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	*M*	9.71**	16.83	22.36	3.18	1.40*	6.13	5.09	3.36
	SD	3.61	7.98	13.63	2.77	1.19	5.05	3.91	1.98
	N	48	48	45	45	45	45	45	45
Low Boys	*M*	8.26	17.48	24.81	3.76	1.88	6.88	5.53	3.77
	SD	3.60	8.21	12.56	2.33	1.55	5.28	3.38	2.09
	N	44	46	43	42	42	42	43	43
High Girls	*M*	8.09***	21.09*	25.17	4.71***	1.52	6.43	6.36	3.60
	SD	4.08	7.86	13.09	3.07	1.35	4.24	3.56	1.03
	N	44	45	42	42	42	42	42	42
Low Girls	*M*	5.92	18.56	22.45	3.04	1.43	6.27	5.57	3.43
	SD	2.66	7.32	10.83	2.41	1.26	4.52	3.34	2.02
	N	52	52	47	47	47	47	47	47

* Significant at the .10 level or less.
** Significant at the .05 level or less.
*** Significant at the .01 level or less.

RESULTS

The performances on the neurotic indices of boys and girls high and low on masculinity are presented in Table 1. The slight variation in the size of the *N*s is due to incomplete protocol on various tests. The results of Table 1 suggest that scoring high or low on the masculinity scale is reflected very little in independent measures of neurotic behavior for boys. The general tendency is for boys low on masculinity to appear more anxious and neurotic, but the differences between them and the high scorers fail to achieve statistical significance. In the case of the girls, the results are more substantial. Girls high on masculinity tend to be more impulsive and anxious than girls low on masculinity, and show more neuroticism about their home life (parental rejection, etc.).

The performances on the neurotic indices of boys and girls high and low on femininity are presented in Table 2. The results indicate

TABLE 2

NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE FEMININITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	*M*	10.33**	18.26**	26.00**	3.87	1.71	7.03*	6.16**	3.61*
	SD	3.57	7.86	13.04	3.02	1.24	4.68	3.64	1.73
	N	43	43	38	38	38	38	38	38
Low Boys	*M*	8.52	14.31	20.73	3.85	1.52	5.27	4.41	2.85
	SD	3.60	6.67	9.99	2.29	1.52	3.87	2.68	1.97
	N	42	42	41	40	40	40	41	41
High Girls	*M*	7.49	19.86	25.25	4.20	1.80	6.65	4.80	3.05
	SD	3.32	7.13	13.41	3.19	1.54	4.58	6.16	2.10
	N	41	43	40	40	40	40	40	40
Low Girls	*M*	7.00	20.73	25.77	4.14	1.73	6.70	4.52	3.70
	SD	2.92	6.79	11.24	2.66	1.19	4.39	5.78	1.89
	N	48	48	44	44	44	44	44	44

* Significant at the .10 level, or less.
** Significant at the .05 level, or less.

TABLE 3

NEUROTIC INDICES OF BOYS AND GIRLS HIGH AND LOW ON THE MASCULINITY SCALE AND IN THE MIDDLE RANGE ON THE FEMININITY SCALE

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	*M*	9.02	16.90	24.30	3.60	1.50	6.15	5.50	3.60
	SD	2.96	8.31	11.36	2.62	1.29	4.52	3.40	1.88
	N	21	21	20	20	20	20	20	20
Low Boys	*M*	9.35	21.40	29.33	4.06	2.06	8.39	6.94	4.56
	SD	2.88	9.06	13.45	2.29	1.47	5.33	3.79	1.99
	N	17	18	18	18	18	18	18	18
High Girls	*M*	7.53	20.60	21.71	3.50	1.64	5.50	5.07	3.93
	SD	4.25	9.36	7.91	2.10	.78	3.31	1.76	1.87
	N	15	14	14	14	14	14	14	14
Low Girls	*M*	5.82	17.78	21.92	3.13	1.17	6.38	5.43	3.38
	SD	2.79	8.08	11.46	2.77	1.18	4.58	3.99	2.21
	N	27	27	24	24	24	24	24	24

that boys who are high on femininity are more impulsive, anxious, and neurotic than boys who are low on femininity, and that their neurotic concerns focus upon physical problems, feelings of insecurity, and irritability. From Table 2 it can be seen that girls high and low on femininity do not differ significantly in their performances on the neurotic indices.

In order to examine the possibility that it was not the variation in masculine or feminine identification but response-set which accounted for some of the differences obtained, a further analysis was undertaken. This analysis was similar to that described above, except that in order to decrease the possibility of response-set being of importance, boys and girls were chosen who were high and low on each scale, but were within the middle range on the other scale (i.e., their scores on the opposite-sex scale were between the 25 and 75 percentiles). Thus, one of the groups of boys and girls was of those who had high or low masculinity scores (upper and lower quartiles), but were within the middle range on femininity scores. Table 3 presents the results on the neurotic indices of boys and girls high and low on masculinity and middle range on femininity. Though none of the means achieve statistical significance, boys low on masculinity and middle on femininity tend to be more anxious and neurotic than boys high on masculinity and middle on femininity, with the focus of concerns on the physical and inferiority subscales of the Brown Personality Inventory. For the girls, those high on masculinity and middle on femininity tend to be more impulsive and anxious, with concerns about physical symptoms.

Table 4 presents the results on the neurotic indices of boys and girls high and low on femininity and middle range on masculinity. There is only one instance in which significant differences are found, but boys high on femininity and middle range on masculinity tend to be more anxious and neurotic, with the focus of concerns about physical, inferiority, and irritability feelings. For girls, those low on femininity and middle on masculinity tend to score higher on measures of anxiety and neuroticism, with concerns surrounding the home. Apparently, both scales are of some use in the evaluation of sex role identification, though this confusion is reflected most clearly by the opposite sex scale. It is noteworthy that though they do not achieve statistical significance (possibly because of the decreased size of the *N*), a number of the differences between the highs and lows in this latter analysis are of some magnitude, and are in the same direction as the scores when response-set is not controlled.

Discussion

The results of the present study show that the opposite-sex scale has a higher relationship to various measures of emotional stability than does the like-sex scale. This suggests that the opposite-sex scales are more effective in the discrimination of faulty sex role identification. The like-sex scales, though discriminating between high and low scorers of the same sex, were not equally effective in detecting faultiness of sex role identification as reflected in these independent measures of emotional stability. Apparently it would be most economical to use the opposite-sex scale in diagnosing sex role identification.

The measures of masculinity and femininity used in this study required children to indicate whether they played or did not play certain games. In addition, children were required to indicate whether they liked or disliked the games that they played. The fact that the responses of both boys and girls to play and game items of their own sex were not significantly related to measures of emotional stability suggests that the role expectations for each sex are fairly explicit in this area of like-sex behavior. Apparently, boys are clear about what boys are supposed to do as boys, and girls are clear about what girls are supposed to do as girls, so that irrespective of sex role confusion, both sexes may make enough conventionally expected responses on the same sex scale for it not to be useful as an indicator of sex role confusion. On the other hand, the greater effectiveness of the opposite-sex scale suggests that role expectations of this sort may be more ambiguous. That is to say, children are less certain about how much interest they should show in the activities of the opposite sex.

Examination of the qualitative differences between boys and girls who are high scorers on the opposite-sex scale and the independent measures of emotional stability shows that there are important sex differences in the type of sex role confusion yielded by this scale. For example, the boys who are high scorers on the femininity scale appear to show considerably more confusion than the girls who are high scorers on the masculinity scale. In previous papers, the authors have shown that femininity in boys is associated with high scores both on anxiety and impulsiveness (Sutton-Smith & Rosenberg, in press-a, in press-b). Those findings are confirmed by the present study. In addition, the fact that the anxiety is expressed mainly on the subscales insecurity, irritability, and physical problems suggests that these boys have a very uncertain "self" picture. On the one hand they are anxious about their personal and physical selves; on the other, they are prepared to indulge in impulsive acting

TABLE 4

Neurotic Indices of Boys and Girls High and Low on the Femininity Scale and in the Middle Range on the Masculinity Scale

Group		Imp	Anx	BPI	Ho	Sch	Phy	Inf	Irr
High Boys	*M*	10.13	18.27**	24.81	3.44	1.75	7.00	5.25	3.25
	SD	3.11	7.40	13.96	3.14	1.44	4.70	2.86	1.39
	N	15	15	16	16	16	16	16	16
Low Boys	*M*	9.50	13.27	20.52	4.19	1.29	5.19	4.19	2.57
	SD	3.42	7.10	9.01	2.72	1.16	3.28	2.58	1.87
	N	22	22	21	21	21	21	21	21
High Girls	*M*	7.20	19.67	25.00	3.62	1.74	6.79	6.79	3.58
	SD	1.91	6.70	12.38	3.22	1.58	3.66	3.87	2.03
	N	20	21	19	19	19	19	19	19
Low Girls	*M*	7.00	21.24	27.19	4.90	1.67	6.67	6.95	3.67
	SD	3.25	7.10	10.48	2.36	.98	4.59	3.19	1.67
	N	21	21	21	21	21	21	21	21

** Significant at the .05 level, or less.

out behavior ("I like to throw snowballs," "I like to chase fire engines"); and in addition, they show a greater than normal preference for feminine play activities and games. While this empirically derived picture is a paradoxical one insofar as impulsiveness and femininity would appear superficially to be contraries, it is similar to a finding of Hartley (1959) based on interview materials that there are some boys who choose such a feminine-impulsiveness as a defense. There is some agreement then that at least one important type of sex role confusion in children of this age group is to be found manifest in this feminine-impulsive syndrome. While the dynamics lying behind this syndrome are not at present apparent, the authors favor the view that impulsiveness is a means of warding off anxiety about sex role deviancy (abnormal identification with feminine role preferences) through pseudomasculine acting out behavior (Sutton-Smith & Rosenberg, in press-a).

The highly masculine girls do not show quite the same inconsistencies in their various scores. Their high scores on the impulsiveness scale (on which males have higher average scores than females) are consistent with their high scores on the masculinity scale. Not only are the girls' choice patterns consistent, they are also more culturally acceptable than the choice patterns of boys scoring high on femininity. It is not unusual for girls to choose masculine type activities in Western culture (Brown, 1958); in fact it is increasingly suitable for them to do so (Rosenberg & Sutton-Smith, 1960; Sutton-Smith & Rosenberg, in press-c). Nevertheless, the presence of high anxiety scores in these girls suggests that their masculine identifications do involve some conflict. Apparently, it is perceived as a conflict between themselves and their parents (higher scores on the home subscale) who are seen as severe and rejecting, rather than a conflict in their own self picture. Possibly, their masculine assertiveness may be a form of defense against parental rejection.

Summary

The present study sought to examine the relative effectiveness of like-sex and opposite-sex scales in discriminating sex role identification. Each subject received scores on like as well as opposite-sex scales, and three independent measures of emotional stability. Apparently, children scoring high or low on a like-sex scale do not differ significantly on measures of emotional stability. Children scoring high on opposite-sex scales tend to be more anxious, impulsive, and neurotic than children low on opposite-sex scales. Explanations of the differing symptoms of such sex deviant boys and girls are offered. From the results of the present study, there is some doubt that like-sex scales, as they are traditionally used, are as effective as heretofore suspected in discriminating sex role identification.

References

Brown, D. G. Sex role development in a changing culture. *Psychol. Bull.*, 1958, **55**, 232–242.

Brown, F. A psychoneurotic inventory for children between nine and fourteen years of age. *J. appl. Psychol.*, 1934, **18**, 566–577.

Castaneda, A., McCandless, B. R., & Palermo, D. S. The children's form of the manifest anxiety scale. *Child Develpm.*, 1956, **27**, 317–326.

Hartley, Ruth. Sex role pressures and the socialization of the male child. *Psychol. Rep.*, 1959, **5**, 457–468.

Hathaway, S. R., & McKinley, J. C. *Manual and booklet for the MMPI.* New York: Psychological Corp., 1943.

Henry, G. W. *Sex variants.* New York: Hoeber, 1948.

Rosenberg, B. G., & Sutton-Smith, B. The measurement of masculinity and femininity in children. *Child Develpm.*, 1959, **30**, 373–380.

Rosenberg, B. G., & Sutton-Smith, B. A revised conception of masculine-feminine differences in play activities. *J. genet. Psychol.*, 1960, **96**, 165–170.

Rosenberg, B. G., Sutton-Smith, B., & Morgan, E. E. Historical changes in the freedom with which children express themselves on personality inventories. *J. genet. Psychol.*, in press.

Strong, E. K. *Vocational interests of men and women.* Palo Alto: Stanford Univer. Press, 1943.

Sutton-Smith, B., & Rosenberg, B. G. A scale to identify impulsive behavior in children. *J. genet. Psychol.*, 1959, **95**, 211–216.

Sutton-Smith, B., & Rosenberg, B. G. Impulsivity and sex preference. *J. genet. Psychol.*, in press. (a)

Sutton-Smith, B., & Rosenberg, B. G. Manifest anxiety and game preference in children. *Child Develpm.*, in press. (b)

Sutton-Smith, B., & Rosenberg, B. G. Sixty years of historical change in the game preference of American children. *J. Amer. Folklore*, in press. (c)

Terman, L. M., & Miles, C. C. *Sex and personality.* New York: McGraw-Hill, 1936.

(Received April 21, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 226–233

TEST ANXIETY AND PERFORMANCE UNDER STRESS[1]

ZANWIL SPERBER
Children's Hospital, Philadelphia

There has been much interest in the effects of anxiety on performance in recent years. In a series of studies Taylor, Spence, and others at Iowa showed that subjects (*Ss*) with a high anxiety level performed better on a variety of simple tasks than *Ss* low in anxiety (Spence & Farber, 1953; Spence & Taylor, 1951; Taylor, 1951; Wenar, 1954). With more complex tasks, anxiety level tended to have the opposite effect, low anxiety *Ss* performing better (Farber & Spence, 1953; Maltzman, Fox, & Morrisett, 1953; Montague, 1953; Ramond, 1953; Taylor & Rechtshaffen, 1959; Taylor & Spence, 1952). These results were explained in terms of Hull's theory of learning (Taylor, 1951; Taylor & Rechtschaffen, 1959; Taylor & Spence, 1952). Child (1954) criticized this theoretical approach for concentrating on the characteristics (simplicity or complexity) of the task while ignoring the category of responses *Ss* learn to make to the cues provided by their own anxiety.

Sarason and his colleagues have presented a more complex theory for explaining the role of anxiety in performance. Their theory predicts different effects from small and large amounts of anxiety, and takes into account the *S*'s response to his own anxiety (Mandler & Sarason, 1952; Sarason, Mandler, & Craighill, 1952; Wiener, 1959). They assume that two kinds of anxiety responses may be aroused by a testing situation: those which are self-centered and ego-defensive, and those which are evoked and learned in the course of task performance and are directed toward task completion.

Small amounts of anxiety are held to improve performance by increasing the *S*'s motivation and strengthening his task relevant responses. Larger amounts of anxiety lead to or strengthen the ego-defensive (R_A) responses. These responses are characteristic of the *S* rather than specific to the task, always available in the response repertoire and readily evoked. Since R_A responses are self-centered rather than task relevant they interfere with performance. Individuals with a high anxiety drive will have many R_A responses in their response repertoire. In contrast, in relation to the total number of responses available, low anxiety *Ss* will tend to make more task relevant anxiety responses. Results obtained with Yale students as *Ss* and using a variety of tasks (e.g., Koh's blocks, maze learning, digit symbol, motor learning, projective material) generally support the theory (Mandler & Sarason, 1952; Mandler & Sarason, 1953; Sarason & Mandler, 1952; Sarason et al., 1952; Wiener, 1959).

Sarason and Gordon (1953) also developed a measure of anxiety specific to the test stress situation to be studied, implicitly recognizing that anxiety need not function as a unitary drive, but might be a potential reaction of the individual which is only elicited under certain

[1] A report of part of a dissertation submitted to the University of Michigan in partial fulfillment of the requirements for the PhD degree. The writer wishes to thank E. Lowell Kelly, Chairman, and members of the doctoral committee for their aid and encouragement.

The cooperation of the United States Air Force was secured by Abraham Carp, Chief of the Motivation and Personality Measurement Division of the Human Resources Research Center (HRRC), now called the Air Force Personnel and Training Research Center. His interest in this research is greatly appreciated. The HRRC detachment at Sampson Air Force Base, commanded by Charles J. Meder, participated actively during the entire process of data collection. The assistance given by the entire group, and especially by Raymond J. Stubblebine and William McMackin, is gratefully acknowledged.

This research, however, was not sponsored by the United States Air Force, and the Air Force is not responsible for this report.

conditions. Sperber (1959) has reported correlations between the scores obtained by Air Force recruits on Sarason's Test Anxiety Questionnaire and their scores on the Taylor Manifest Anxiety scale (1953) and Winne's Neuroticism scale (1951). Winne's interpretation of his results indicates that his scale can be considered a general measure of manifest anxiety (pp. 120–121). The highest correlation between the test anxiety and general anxiety measures was +.35, suggesting that a general anxiety measure might not be an adequate gauge of an *S*'s proneness to anxiety in a test situation. Since the interest of the present research was in studying how anxiety, when evoked, operates to affect test performance, Sarason's measure of anxiety proneness specific to testing situations was used.

The purpose of the present study was to determine whether there is any difference in performance between: (*a*) high test anxiety *Ss* under high vs. low stress, (*b*) low test anxiety *Ss* under high vs. low stress, (*c*) high test anxiety *Ss* vs. low test anxiety *Ss* under high stress, and (*d*) high test anxiety *Ss* vs. low test anxiety *Ss* under low stress.

A further purpose was to bring additional evidence to bear on the different theoretical approaches to the problem of anxiety and performance presented by Taylor and Spence (1952) and by Sarason et al. (1952).

Method

Stress Environment

A situation is stressful by virtue of its ability to induce anxiety in individuals subjected to the situation. Lazarus, Deese, and Osler (1952) observed that "The principal problem in the study of behavior under stress has been the production of realistic stress situations" (p. 295). This experiment took place in association with a naturally stressful testing situation. The experimental data were collected at Sampson Air Force Base in December 1953, using new recruits as *Ss*. Prior to testing, an officer lectured them on the importance of the aptitude testing program they were to undergo. Many of the recruits hoped to benefit from the opportunity for some specialized vocational training which would equip them for a post service career; all were interested in getting a good assignment for their service period. The officer told them, in essence, that the realization of their hopes would be determined by the results of the aptitude testing.

The recruits were then tested on two successive days. The first day was devoted to the regular aptitude testing program. The second day was scheduled for HRRC experimental testing, but the recruits were only told to report to the same building for more testing. The use of identical administration procedures by the HRRC and aptitude testing staffs reinforced the *Ss*' tendency to view both days' testing as serving the same purpose.

High stress. *Ss* were allowed to continue believing that the experimental tests, presented as the Abilities Test Battery and administered by uniformed military personnel,[2] were part of the official assessment program. The battery consisted of a large number of timed subtests. The instructions implied that the *Ss* were expected to finish all of the items in each subtest, incompleted items counting as failures. The time limits were so chosen that the *Ss* could not possibly complete all of the items. The very first "test" was an interesting puzzle chosen to challenge and involve the *S*, but not solvable in the time allotted (Cowen, 1952, p. 514). It was included solely for the purpose of starting the *Ss* with a failure experience.

After completing the battery, a 40-minute interlude was devoted to procedures presented as not related to the "assessment program," and administered by a civilian. Then the sergeants announced: "We're going to repeat two of the tests you took this morning. You've had practice on them and you should improve. Let's see how much better you do this time." The first test was the puzzle, which again provided a failure experience. *Ss* then repeated the Letter Substitution Test.

Low stress. The instructions were designed to make it clear that the Abilities Test Battery was experimental, and not part of the regular assessment program, but also to indicate the importance of the tests to the Air Force and to maintain a high level of task oriented motivation for good performance. For these *Ss* the first test was a simple line tracing task which started them with a success experience. The rest of the battery was the same as for the high stress group. The low stress *Ss* were interrupted the same number of times and in the same places, but the instructions allowed them to treat unanswered items as evidence that the test required changes, rather than as signs of personal inadequacy.

After the 40-minute interlude, instead of receiving instructions which implied that an improvement in performance was expected, the low stress *Ss* were told: "We're going to repeat two of the tests you took this morning. You've had practice on them which should help, but you're somewhat tired by now—so let's see what happens. Do the best you can."

To help underscore the dissociation of the tests from the regular assessment program, the test administrator wore civilian clothes. To emphasize that the Air Force really had an interest in the procedure, the proctor, who was in a less active role, remained in uniform.

---

[2] Stubblebine administered the performance tests under both high and low stress conditions, and Swenson proctored all of the testing.

Performance Battery [3]

A series of paper and pencil tests, suitable for group administration, and tapping a range of cognitive functions was used. With the exception of the atmosphere reinforcing puzzle or tracing task, high and low stress *Ss* worked on the same tasks, which are described in the order they appeared in the battery.

Letter series. An expanded version of Thurstone's (1943) PMA Reasoning subtest, containing the original 30 items and 20 new ones constructed by the writer, was administered in two consecutive parts. For each part, containing a random selection of 15 original and 10 new items, 4 minutes were allowed, the parts being scored separately.

Letter substitution. The items were taken from Lazarus and Eriksen's (1952) expanded version of the Form I Digit Symbol test. The *Ss* were allowed $3\frac{1}{2}$ minutes for 200 items. This test was repeated in the post interlude battery.

Revised Minnesota Paper Form Board. The series MB version of this test (Likert & Quasha, 1948), which measures the *S*'s ability to perceive spatial relations, was administered in the usual manner.

Number matching. This test is similar to the number comparison sections of the Minnesota Vocational Test for Clerical Workers (Andrew, Patterson, & Longstaff, 1933), but new items were constructed by the writer. It measures the *S*'s ability to concentrate and attend to detail, the items requiring the *S* to make rapid discriminations of small differences. Each page contained 100 items and was timed and scored separately. There were three consecutive parts to the test, with $4\frac{1}{2}$ minutes allowed per page.

Test Reaction Questionnaire (TRQ)

After the performance testing, the experimental purpose of the tests was explained to the *Ss*. Then a 22-item questionnaire, constructed by the writer and similar in format to the Test Anxiety Questionnaire (Sarason & Mandler, 1952), was administered to ascertain the *S*'s reactions to various aspects of the testing situation.[4] The *S* put a mark on a 16-centimeter line to indicate the strength of his reaction, the ends of the line representing polar responses to the question. Each question was scored by measuring the distance in centimeters from the end of the line to the *S*'s mark.

Personality measures

A slightly modified version of Sarason's Test Anxiety Questionnaire (TAQ) [5] was administered to the *Ss* after the TRQ. Following Sarason and Gordon's (1953) suggestion, local norms, based on the responses of both high and low stress *Ss*, were used in scoring each question. The resulting distribution of scores was very similar to that reported for Yale students (Sperber, 1959).

Other Procedures

The Minnesota Multiphasic Personality Inventory (MMPI) was administered in an afternoon session. Also available for each *S* was his stanine score on the Technician's Specialty Index (TSI) (Dailey, Lecznar, & Brokaw, 1948), considered to be the military test which best measures general intelligence level.[6] For most *Ss*, raw scores on the Armed Forces Qualification Test (AFQT) (Bolanovich, Mundy, Burke, & Falk, 1953), another gauge of intelligence level, were also available.

Subjects

All 399 new recruits who arrived at Sampson AFB to begin basic training the week before this experiment was conducted constituted the original sample tested. The *Ss* were organized in six "flights" with about 65 men in each. Three flights were arbitrarily assigned to each of the stress conditions. Tests were administered to one or two flights at a time, 201 men performing under high and 198 under low stress.

Recruits with a TSI in the lowest two stanines, or an AFQT score less than 15 were considered too low in intelligence to participate, and their data were not analyzed. Twenty percent were eliminated by these criteria. *Ss* whose MMPI records did not meet acceptable standards on the validity checks (Hathaway & Meehl, 1951) were also dropped, leaving 146 high stress and 148 low stress *Ss*. These men supplied the data on which were based the norms for scoring the TAQ, the correlations between test anxiety and general anxiety, and the analyses of reactions to the experimental conditions (TRQ responses).

The performance of only those *Ss* who scored in the highest and lowest quartiles on the TAQ was studied. High test anxiety (HTA) *Ss* had scores of 24 or higher, low test anxiety (LTA) *Ss* scored 11 or less. There were 61 *Ss* under high stress, 32 HTA and 29 LTA. Under low stress there were 71 *Ss*, 33 HTA and 38 LTA. These four groups were well matched with respect to age, years of education, and intelligence. For each group the mean age was close to 19, mean years of education was about 11, and the mean TSI stanine was approximately 5.5.

Results

Effectiveness of the Experimental Operations

The distribution of responses of the 146 high stress and 148 low stress *Ss* to each of the questions of the TRQ were compared by

---

[3] I wish to thank Robert H. Bauernfeind, Editor of the Test Department, Science Research Associates, for permission to use the Letter Series items from the PMA (Thurstone, 1943), and Richard S. Lazarus for making available the extended version of the Digit Symbol test.

[4] Copies of the TRQ may be obtained from the writer.

[5] I wish to thank Seymour B. Sarason for making available a copy of his scale, and granting permission for such changes in wording as were necessary to adapt it for use in a different context.

[6] Personal communication from Abraham Carp.

means of the Marshall test (Smith, 1953). The high stress *Ss* differed significantly from the low stress *Ss* in being convinced that the testing would decide their Air Force assignments. The low stress *Ss* more strongly accepted the idea that testing was for experimental purposes. They differed significantly ($p < .05$) on 11 of the 22 questions. High stress *Ss* felt more anxious, a greater pressure to do well, and a greater degree of unpleasantness in the testing situation. The two groups did not differ in their evaluation of how much their emotions and anxieties had interfered with performance.

The differences in TRQ responses of *Ss* scoring above and below the median on the test anxiety measure were tested separately for *Ss* who had performed under high stress and *Ss* who had performed under low stress. Under high stress, HTA and LTA *Ss* responded differently to 10 of the 22 questions (Marshall test, $p < .05$). HTA *Ss* more strongly believed that their Air Force assignment depended on their performance, felt more anxious, thought their emotions influenced their performance adversely, and considered the testing unpleasant. The groups did not differ with respect to test motivation.

Under low stress, HTA and LTA *Ss* responded differently to 4 out of 22 questions. There was a stronger tendency for HTA *Ss* to believe that their Air Force assignment might be at stake, and they responded with more anxiety to various aspects of the situation. There were no differences in motivation, feelings about the pleasantness of the situation, or evaluation of the influence their emotions had on performance.

TABLE 1

PERFORMANCE DIFFERENCES OF MATCHED HIGH TEST ANXIETY SUBJECTS UNDER HIGH VS. LOW STRESS

Test	High Stress Mean	High Stress σ	Low Stress Mean	Low Stress σ	t^a
Letter Series 1	4.5	2.0	5.7	2.7	1.58
Letter Series 2	7.1	3.6	7.0	2.8	.09
Letter Substitution A	103.4	21.5	98.7	13.5	.83[b]
Paper Form Board	39.6	8.8	40.6	8.0	.39
Number Matching 1	60.3	11.4	53.0	10.8	2.09*
Number Matching 2	61.4	10.2	52.9	10.9	2.54*
Number Matching 3	60.5	11.7	54.7	14.6	1.39
Letter Substitution B	133.5	27.9	129.3	17.9	.56[c]
Intelligence (TSI)	5.4	1.5	6.0	1.5	1.22

[a] $df = 40$.
[b] High Stress *Ss* were more variable ($F = 2.53$, $p < .05$).
[c] High Stress *Ss* were more variable ($F = 2.43$, $p < .05$).
* $p < .05$.

TABLE 2

PERFORMANCE DIFFERENCES OF LOW TEST ANXIETY SUBJECTS UNDER HIGH VS. LOW STRESS

Test	High Stress Mean	High Stress σ	Low Stress Mean	Low Stress σ	t^a
Letter Series 1	4.9	2.2	6.3	2.4	2.17*
Letter Series 2	6.2	2.8	7.9	3.0	2.42*
Letter Substitution A	98.0	19.0	108.9	29.2	1.84[b]
Paper Form Board	39.6	8.5	39.8	6.3	.13[c]
Number Matching 1	55.6	11.9	58.4	13.6	.90
Number Matching 2	55.9	10.4	57.0	12.8	.40
Number Matching 3	55.9	13.0	62.0	16.4	1.65
Letter Substitution B	133.5	23.3	137.6	33.2	.57[d]
Intelligence (TSI)	5.2	1.5	5.7	1.6	1.34

[a] $df = 65$.
[b] Low Stress *Ss* were more variable ($F = 2.38$, $p < .02$); t and the value of t necessary for significance were calculated using the method described by Edwards (1950, p. 167–169).
[c] High Stress *Ss* were more variable ($F = 1.83$, $p < .05$).
[d] Low Stress *Ss* were more variable ($F = 2.03$, $p < .05$).
* $p < .05$.

Performance

The 32 HTA *Ss* tested under high stress performed significantly better than the 33 HTA *Ss* working under low stress on the three Number Matching tests. Critical ratios of the mean differences yielded p values of $< .01$, $< .001$, $< .05$. These results, however, were based on groups differing not only with respect to stress treatments, but also in test anxiety level. The high stress HTA *Ss* were significantly more anxious than the low stress HTA *Ss* (Marshal test, $p = .05$). To control for this the HTA groups were matched for level of test anxiety by a random selection procedure (Sperber, 1956, p. 136), 21 HTA *Ss* remaining in each stress group.

The results of the t tests of the differences in mean performance of the matched HTA groups working under high vs. low stress is given in Table 1. The groups also were adequately matched for intelligence. Stress clearly influenced the performance of HTA *Ss*, the groups tested under high stress performing significantly better on two tests, Number Matching 1 and 2.

The results of the t tests of the differences in mean performance of the LTA *Ss* tested under high vs. low stress are presented in Table 2. The groups were adequately matched both for intelligence and level of test anxiety. The LTA *Ss* consistently had higher mean scores under low stress. For both reasoning

tests the mean differences reached statistical significance.

The *t* tests of differences in mean performance of HTA and LTA *Ss*, both tested under high stress are presented in Table 3. On six of the eight measures the HTA performed better. The only statistically significant difference was on Number Matching 2, where the HTA group was superior in performance. The results on the other two Number Matching tests supported this finding, both the mean differences favoring the HTA *Ss*, and approaching significance. For Number Matching 1 the value of *p* was .08, and for Number Matching 3 it was .10.

The critical ratios, testing the significance of the differences in mean level of performance under low stress of HTA and LTA *Ss*, are reported in Table 4. On all eight tests the LTA *Ss* were superior in performance. Two of the mean differences were statistically significant. For three more tests, namely, Number Matching 1 and 2, and Letter Substitution B, the differences were nearly significant, the values of *p* reaching .06, .09, and .07, respectively.

Discussion

Our findings do not appear to be compatible either with the Iowa or with the Sarason theory. With respect to the Iowa position, we find that under high stress, HTA *Ss* perform better on Number Matching than do LTA *Ss*. Here higher drive seems to facilitiate performance, the correct responses for this task presumably being uppermost in the average

TABLE 3

Performance Differences of High Test Anxiety and Low Test Anxiety Subjects under High Stress

Test	HTA Mean	HTA σ	LTA Mean	LTA σ	t
Letter Series 1	4.6	2.0	4.9	2.2	.43
Letter Series 2	7.2	3.7	6.2	2.8	1.11
Letter Substitution A	106.0	22.8	98.0	19.0	1.46
Paper Form Board	37.5	12.6	39.6	8.5	.72[b]
Number Matching 1	61.0	11.0	55.6	11.9	1.81
Number Matching 2	62.1	10.2	55.9	10.4	2.31*
Number Matching 3	61.3	11.7	55.9	13.0	1.69
Letter Substitution B	134.8	28.9	133.5	23.3	.18
Intelligence (TSI)	5.4	1.5	5.2	1.5	.54

[a] $df = 59$.
[b] HTS *Ss* were more variable ($F = 2.19$, $p < .05$).
* $p < .05$.

TABLE 4

Critical Ratio of Performance Differences of High Test Anxiety and Low Test Anxiety Subjects under Low Stress

Test	HTA Mean	HTA σ	LTA Mean	LTA σ	*CR*
Letter Series 1	5.8	2.8	6.3	2.4	.81
Letter Series 2	7.1	3.1	7.9	3.0	1.15
Letter Substitution A	96.6	17.7	108.9	29.2	2.18[a]*
Paper Form Board	39.2	7.3	39.8	6.3	.48
Number Matching 1	52.9	10.9	58.4	13.6	1.90
Number Matching 2	51.9	12.1	57.0	12.8	1.73
Number Matching 3	54.1	14.3	62.0	16.4	2.15*
Letter Substitution B	125.4	18.3	137.6	33.2	1.95[b]
Intelligence (TSI)	5.5	1.6	5.7	1.6	.43

[a] LTA *Ss* were more variable ($F = 2.73$, $p < .01$); *CR* necessary for significance at $p < .05$ with the observed heterogeneous variances is 2.03 (Edwards, 1950, p. 167–169).
[b] LTA *Ss* were more variable ($F = 3.28$, $p < .001$).
* $p < .05$.

S's response hierarchy. On the same tests under low stress, however, LTA *Ss* perform better than HTA *Ss*, higher drive level apparently interfering with performance. Since the same test is involved in both cases, we can not readily attribute the reversal of the effects of a heightened anxiety drive to the different types of response hierarchies elicited by simple as contrasted with complex tasks.

Taylor (1956) has commented on the many characteristics other than drive level in which anxious and nonanxious *Ss* may differ, and which may influence performance (p. 303). The interest of the Iowa group in their study of anxiety has been restricted to the effects of drive, although the influence of anxiety drive per se on performance has been acknowledged to be small (Spence & Taylor, 1953; Taylor, 1956). Our own results confirm Taylor's suggestion that to understand the performance of anxious and nonanxious *Ss*, a broader approach to their characteristics is necessary. Such an approach will be indicated below.

The effects of anxiety on the performance of our Air Force recruits were consistently opposite to the results reported by Sarason. For the Yale *Ss*, strong anxiety—as manifested by *Ss* who were both high in test anxiety and tested under high stress—interfered with performance, but for the recruits studied in the present research strong anxiety improved performance. For the Yale students, moderate anxiety—as manifested by HTA *Ss* tested under low stress, or LTA *Ss* studied

under high stress—facilitated performance, but for our recruits it was associated with poorer performance. For the Yale students, a near lack of anxiety—such as would be the case for LTA *Ss* studied under low stress—was paralleled by a lowered level of performance. In contrast, for our Air Force recruits this lack of anxiety was associated with an improvement in performance.

A rationale for the performance of our recruits, and the difference between their performance pattern and that of Yale students can be offered in terms of the *S*'s assessment of the personal importance of the testing situation, and the tendency to use avoidance or vigilance as a defense against anxiety, depending on the importance of a situation.

Douvan (1956) has demonstrated that an *S*'s evaluation of the importance of a testing situation will depend on experiences associated with his social status. She showed that working class adolescents become involved and strive to do well in a situation where good performance can earn them an immediate reward, but are not motivated to do well when no direct reward is at stake. In contrast, the middle class adolescent appears to be strongly motivated for successful performance in both situations. Our *Ss* were predominantly from the working class (Sperber, 1959). McNeil (1953) has shown that a low evaluation of formal education is characteristic of families from a working class background. Although the recruits were about the same age as Yale students, a majority of them never completed high school. Their fewer years of education is assumed not to be simply a function of lower intelligence level, but also to reflect skepticism about the value of an academic education. This attitude was apparently shared by their parents since many of them were young enough to require parental permission to enlist. The Yale *Ss* were predominantly from the middle class (Sarason & Mandler, 1952) where characteristically there are strong pressures for academic achievement (McNeil, 1953).

The operation of vigilance and avoidance as defenses against anxiety is conceived in terms akin to Bruner and Postman's (1947) description of perceptual vigilance and defense:

> In any given situation the organism singles out what it considers to be the environment's most relevant aspects—relevant to adaptation in the situation. So long as the situation is not *too* threatening or *too* exacting, avoidance of meaning may be emotionally the most economical response. But in situations which are highly threatening and highly exacting, the most adaptive perceptual response is frequently the one which takes most vigilant account of "reality" (p. 76).

With reference to the present findings, we view the high stress situation, testing which the *S* thought would determine his future, as one in which the *S* would be motivated to participate to the best of his ability. Participative rather than avoidance behavior seemed most in the *S*'s self-interest, and we assume both HTA and LTA *Ss* were involved in the performance situation. We attribute the superior performance of the HTA *Ss* to the function of anxiety as an internal stimulus which could constantly remind them of both the importance and danger of the situation, and reinforce their motivation to be alert and perform well. The performance of the LTA group did not benefit from this increase in vigilance. This rationale seems especially plausible when we consider the nature of the Number Matching tests, on which the significant performance differences occurred. The greater anxiety of the HTA *Ss* would maintain their motivation to perform well on a repetitive, boring task, given at the end of a long session. An increment in perceptual vigilance would improve their discrimination of small differences.

Under low stress the LTA group performed better. The TRQ data allow us to assume that for the HTA *Ss* the situation was sufficiently like one involving real tests, to elicit feelings of anxiety. Since the instructions used assured the *Ss* that it was safe to be unconcerned about one's own performance, i.e., no immediate gain for them was possible, they would tend to defend themselves by withdrawal when the task became noxious. This need to avoid became strong enough to affect performance only on the more repetitious tasks used.

Our analysis of the results of the present study has been in terms which are similar in many respects to the theoretical formulations advanced by Lazarus and his colleagues (e.g., Vogel, Baker, & Lazarus, 1958) and by Rue-

bush (1960). The present study and other recent researches (Ruebush, 1960; Vogel et al., 1958; Wiener, 1959) have all presented results which point up the complex interaction of anxiety, motive, defense, and task variables. A prediction of whether anxiety leads to an improvement or decrement in the performance of a given *S* appears to depend on the answers to four related questions: (*a*) Does the *S* accept the performance situation as something so important that he must participate in it, or as something that can appropriately be avoided? (*b*) Does the anxiety aroused in the *S* by the situation lead to motivations which are more important as determinants of his behavior than the actual amount of anxiety elicited? (*c*) With respect to the direct effect of anxiety on performance, what kind(s) of defensive behavior is (are) engendered in the *S* by varying amounts of anxiety? (*d*) What relationship obtains between the nature of the *S*'s defenses and the structure and demands of the task?

Summary

Air Force recruits, scoring in the highest and lowest quartiles on Sarason's measure of test anxiety were tested under high and low stress. The significant performance differences were:

1. Under high stress, HTA *Ss* performed better than LTA *Ss* on the Number Matching tests.

2. Under low stress, LTA *Ss* performed better than HTA *Ss* on the Letter Substitution and Number Matching tests.

3. Performance of HTA *Ss* under high stress was superior to that of a group matched on test anxiety but subject to low stress on two of the three Number Matching tests.

4. Performance of LTA *Ss* under low stress was superior to that of a group matched on test anxiety but subject to high stress on the two Letter Series tests.

The results were discussed in relation to the findings of Taylor and Spence, and those of Sarason et al., and a theoretical formulation advanced which emphasized the complex interaction of anxiety, motive, defense, and task variables.

References

Andrew, D. M., Patterson, D. C., & Longstaff, H. P. *Minnesota Clerical Test*. New York: Psychological Corp., 1933.

Bolanovich, D., Mundy, J. P., Burke, L. K., & Falk, G. H. Development of the Armed Forces Qualification Test, Forms 3 and 4. *USA TAGO Personnel Res. Br. tech. res. Rep.*, 1953, No. 1078.

Bruner, J. S., & Postman, L. Emotional selectivity in perception and reaction. *J. Pers.*, 1947, **16**, 69–77.

Child, I. L. Personality. *Annu. Rev. Psychol.*, 1954, **5**, 149–170.

Cowen, E. L. The influence of varying degrees of psychological stress on problem-solving rigidity. *J. abnorm. soc. Psychol.*, 1952, **47**, 512–519.

Dailey, J. T., Lecznar, W. B., & Brokaw, L. D. Development of the Airman Classification Test Battery. *Hum. Resources Res. Cent. res. Bull.*, 1948, No. 48-5.

Douvan, Elizabeth. Social status and success strivings. *J. abnorm. soc. Psychol.*, 1956, **52**, 219–223.

Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.

Farber, I. E., & Spence, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 120–125.

Hathaway, S. R., & Meehl, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: Univer. Minnesota Press, 1951.

Lazarus, R. S., Deese, J., & Osler, Sonia F. The effects of psychological stress upon performance. *Psychol. Bull.*, 1952, **49**, 293–317.

Lazarus, R. S., & Eriksen, C. W. Effects of failure stress upon skilled performance. *J. exp. Psychol.*, 1952, **43**, 100–105.

Likert, R., & Quasha, W. H. *The revised Minnesota Paper Form Board Test: Manual*. New York: Psychological Corp., 1948.

McNeil, E. B. Conceptual and motoric expressiveness in two social classes. Unpublished doctoral dissertation, University of Michigan, 1953.

Maltzman, I., Fox, J., & Morrisett, L. Some effects of manifest anxiety on mental set. *J. exp. Psychol.*, 1953, **46**, 50–54.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 166–173.

Mandler, G., & Sarason, S. B. The effect of prior experience and subjective failure on the evocation of test anxiety. *J. Pers.*, 1953, **21**, 336–341.

Montague, E. K. The role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, **45**, 91–96.

Ramond, C. K. Anxiety and task as determiners of verbal performance. *J. exp. Psychol.*, 1953, **46**, 120–124.

Ruebush, B. K. Interfering and facilitating effects of test anxiety. *J. abnorm. soc. Psychol.*, 1960, **60**, 205–212.

Sarason, S. B., & Gordon, E. M. The Test Anxiety Questionnaire: Scoring norms. *J. abnorm. soc. Psychol.*, 1953, **48**, 447–448.

SARASON, S. B., & MANDLER, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, **47**, 810–817.

SARASON, S. B., MANDLER, G., & CRAIGHILL, P. G. The effect of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47**, 561–565.

SMITH, K. Distribution-free statistical methods and the concept of power efficiency. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences*. New York: Dryden, 1953.

SPENCE, K. W., & FARBER, I. E. Conditioning and extinction as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 116–119.

SPENCE, K. W., & TAYLOR, JANET A. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *J. exp. Psychol.*, 1951, **42**, 183–188.

SPENCE, K. W., & TAYLOR, JANET A. The relation of conditioned response strength to anxiety in normal, neurotic, and psychotic subjects. *J. exp. Psychol.*, 1953, **45**, 265–272.

SPERBER, Z. A study of the role of anxiety level and defense preference in performance under stress. Unpublished doctoral dissertation, University of Michigan, 1956.

SPERBER, Z. The Test Anxiety Questionnaire: Scoring norms for a noncollege population. *J. abnorm. soc. Psychol.*, 1959, **58**, 129–131.

TAYLOR, JANET A. The relationship of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, **41**, 81–92.

TAYLOR, JANET A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

TAYLOR, JANET A. Drive theory and manifest anxiety. *Psychol. Bull.*, 1956, **53**, 303–320.

TAYLOR, JANET A., & RECHTSCHAFFEN, A. Manifest anxiety and reversed alphabet printing. *J. abnorm. soc. Psychol.*, 1959, **58**, 221–224.

TAYLOR, JANET A., & SPENCE, K. W. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, **44**, 61–64.

THURSTONE, L. L. *Primary Mental Abilities Test*. Chicago: Science Research Associates, 1943.

VOGEL, W., BAKER, R. W., & LAZARUS, R. S. The role of motivation in psychological stress. *J. abnorm. soc. Psychol.*, 1958, **56**, 105–112.

WENAR, C. Reaction time as a function of manifest anxiety and stimulus intensity. *J. abnorm. soc. Psychol.*, 1954, **49**, 335–340.

WIENER, G. The interaction among anxiety, stress instructions, and difficulty. *J. consult. Psychol.*, 1959, **23**, 324–328.

WINNE, J. F. A scale of neuroticism: An adaptation of the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1951, **7**, 117–122.

(Received April 25, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 234–238

WECHSLER'S DETERIORATION RATIO IN CLINICAL PRACTICE

T. G. CROOKES
St. John's Hospital, Aylesbury, England

Studies on Wechsler's Deterioration Ratio have mostly suggested that it is of very limited value in diagnosing brain damage. None of those, for instance, quoted by Yates (1954) showed any very convincing differentiation of brain damaged from other groups. Even where it distinguished brain damaged from "normals," it did not show any appreciable difference between brain damaged and other psychiatric groups (Hall, 1952; Rogers, 1950), and it is this latter distinction which it is of practical value to make. The usual method is to take groups of known diagnoses, apply the tests to them, and compare the results. For reasons given below, this was not thought to be the most satisfactory way in this case, and it seemed to be of interest to examine the ratios in a group referred for routine purposes over a period of time, to look into the distribution of the values and compare them with the final diagnoses.

The actual ratio used is not exactly the same as Wechsler's. The same four "don't hold" tests are used, but only two "hold" tests: Vocabulary and Picture Completion. The score on these is doubled, and the percentage loss is calculated in the usual way with Wechsler's age allowances. Object Assembly was omitted because it does not seem to be a good hold test even on Wechsler's data (1944, p. 150), and this is a rather embarrassing test for adults, except the duller ones. Information was omitted to keep the balance of verbal and performance tests. The ratio will be referred to as DR.

Procedure

Wechsler's test is applied routinely to almost all patients referred for psychological examination at the above hospital. It is a regional National Health Service hospital, catering for all kinds of mental illness.

The DR was calculated for all male inpatients referred over a period of 6 years, whatever the reason for referral, providing they had done all the tests necessary for the ratio. In all cases it was the Wechsler-Bellevue Scale, Form I, and where the test had been repeated, the figures of the first testing were used. The diagnoses were the final diagnoses made on discharge or death, and, where the patient is still here, the latest diagnosis at the time of writing supplied by the psychiatrist in charge of the case.

There are altogether 261 men, of whom 171 were less than 50 years of age at the time of testing, and 90 were 50 or over. The main comparison is between those in whom the diagnosis implies physical interference with the brain (the "organic" group) and the rest. The under-50 organic group consists of 23 patients, 12 epileptics, 3 head injuries, 3 toxic conditions (2 of them alcoholic), 2 cerebral syphilis, 1 confusional state in disseminated sclerosis, and 2 cases of unspecified brain disease. The over-50 organic group contains 22 patients, 12 with senile or presenile dementia, 1 alcoholic dementia, 5 toxic conditions, 2 cerebral syphilis, 1 head injury, and 1 cerebrovascular disease.

The remainder were divided into five broad diagnostic categories, to see if any type of score was peculiar to any group, and to see if the results agreed with the finding of Rogers (1950) that all maladjusted groups showed higher DRs than normals, but did not differ among themselves. The under-50 "nonorganic" group consists of 30 depressives (including manics), 45 schizophrenics, 29 neurotics, 33 psychopaths, and 11 paranoid states (including paraphrenics). The over-50 nonorganic group contains 40 depressives, 1 schizophrenic, 10 neurotics, 9 psychopaths, and 8 paranoid states.

Results

Table 1 shows the means and ranges of DRs for the organic and nonorganic groups.

In both age groups the organics have considerably higher values than the others. (Comparing organic and nonorganic means, for the under-50 groups, $t = 4.866$, $p < .001$; for the over-50, $t = 4.596$, $p < .001$.) But for practical purposes, one needs to know to what extent they can be differentiated by taking

TABLE 1

MEAN DETERIORATION RATIO, ORGANIC AND NONORGANIC

Group	Under-50			Over-50		
	N	*M*	Range	*N*	*M*	Range
Organic	23	+28.9	+59 to −22	22	+23.0	+57 to −18
Nonorganic	148	+ 8.8	+57 to −42	68	− 0.4	+42 to −52

some critical value or "cutting score." Table 2 gives the number falling above +20 (Wechsler's suggested critical score) and above +30, by age group, and whether organic or not.

If the two age groups are combined, we find using a score of 20 that 76.6% are correctly classified (organics 64.4% and nonorganics 79.2%). Using the score of 30, 85.1% are correctly classified (organics 53.3% and nonorganics 91.7%). The increase in total percentage correct obtained by raising the critical score from 20 to 30 is mainly due to the fact that the nonorganic group is much larger than the organic; by raising the score the discrimination of the nonorganic group is improved, that of the organic group is made worse, and as the former is so much bigger they contribute more to the total percentage. However, it will be noted that with the score of 30, the mean of the organic and nonorganic percentages (72.5) is slightly greater than their mean with the score of 20 (71.8). This mean can be considered as the percentage correct if the two groups are made equal in number; so the cutting score of 30 appears slightly better. In any case, it is probably of more value in this kind of problem to have a more extreme score, which nearly eliminates one group, providing it leaves a substantial number of the other. For instance, a score which gave 100% nonorganics correct and 50% organics, would be of more practical value than one which gave 75% of each, although the total percentage correct is the same. In the first case you would be able to feel virtual certainty when some scores came up, but never in the second case. The whole general question of how proportions obtained in this way can be used in a practical situation, where the size of the populations from which you are drawing your samples is unknown, is considered in the discussion.

TABLE 2

PROPORTIONS DISTINGUISHED BY CUTTING SCORES OF +20 AND +30

Group	Under-50		Over-50	
	DR <21	DR >20	DR <21	DR >20
Organic	7	16	9	13
Nonorganic	112	36	59	9
	(75% correct)		(80% correct)	
	DR <31	DR >30	DR <31	DR >30
Organic	10	13	11	11
Nonorganic	134	14	64	4
	(86% correct)		(83% correct)	

Table 3 shows the mean values of the DRs for the nonorganic patients divided into five diagnostic categories, again separated into the two age groups.

The most striking thing about Table 3 is that the younger patients consistently give higher values than the older ones, while within each group, there is little difference between diagnoses. This is fairly clear by inspection, but a simple analysis of variance was carried out for each age group (leaving out the schizophrenic in the over-50). In each case, the between diagnoses mean square was smaller than that within diagnoses. The schizophrenics and paranoid states, under-50, were combined (one can find reasons why they might be combined) and compared with the rest of their age group. This gives $t = 1.31$, $p > .1$. In the older age group, the neurotics were compared with the rest; $t = 1.63$, $p > .1$. It will be seen also that the high values of the DR are fairly evenly distributed among the various groups.

From the way in which the DR is calculated, the expected mean value for any age group is nought, if the group is comparable with Wechsler's standardization group. For the over-50 group, the mean is very close to

TABLE 3

DETERIORATION RATIO OF NONORGANIC PATIENTS BY DIAGNOSIS

Group	Under-50				Over-50			
	N	*M*	Range	$n > 30$	*N*	*M*	Range	$n > 30$
Depressives	30	+ 6.8	+47 to −42	3	40	+0.8	+42 to −52	3
Schizophrenics	45	+10.4	+57 to −34	5	1	—	+14	0
Paranoid states	11	+14.1	+31 to −14	1	8	+0.1	+26 to −36	0
Neurotics	29	+ 8.1	+46 to −26	4	10	−9.6	+32 to −29	1
Psychopaths	33	+ 7.2	+49 to −29	1	9	+2.6	+28 to −16	0
Total	148	+ 8.8	+57 to −42	14	68	−0.4	+42 to −52	4

nought, but not for the under-50 group; comparing their mean (+8.81) with its standard error (1.50) gives $t = 5.87$, $p < .001$.

DISCUSSION

In differentiating organics from nonorganics, the results here seem rather better than those obtained in most studies. One reason for this is probably that these are not selected groups of patients, but patients referred for information in the routine way. This means that in most cases the diagnosis was in doubt at the time of referral, and the deterioration was not marked. Gutman (1950), in summing up her study, says that hers were definite cases and that in less advanced cases even poorer differentiation would be expected. In general, the principle of taking extreme cases is excellent, but in this case it is not appropriate. The theory behind the ratio is that certain types of activity deteriorate more rapidly than others with aging and other forms of damage to the brain. It is not suggested that the other activities do not deteriorate at all. Clearly with gross dementia all abilities, including Vocabulary, disappear, cf. Yates (1956). It might be expected that the discrepancy would be most clear in the early stages, while later on all abilities declined to a common level.

Hall's (1952) organic cases are described as mostly moderately deteriorated, and he found only 5 out of 24 with ratios over +29 and only 7 over +19. His inclusion of Object Assembly may partly account for this. The group described by McFie and Piercy (1952), consisting of patients with localized cerebral lesions, gave results fairly close to those of the present study. Of 56 DRs calculated, they had 24 (43%) over +29 and 37 (66%) over +19. For my whole organic group, the equivalent percentages are 53 and 67.

An interesting problem is the question of what weight can be given to an actual DR in a practical situation. If one has obtained a ratio greater than 30, what can be said about it? In this group, 18 out of 216 nonorganics obtained such a score, or 1 in 12; 24 out of 45 organics did so, rather more than half. So, in a sense, it can be said that such a score has a good probability of being organic; an organic case is much more likely to produce one than a nonorganic. However, if one considers the distribution of the 42 such scores obtained, only 24 are organic, and the probability of a given score of over 30 being organic is only 4/7, not much more than a half. In general, this is because the probability of an event's belonging to one of two categories depends not only on the relative frequency of the event within each category, but also on the relative frequency of the two categories in the population from which the event is drawn. This seems to be a general difficulty in all tests of this kind except where a certain score is found in virtually 100% of one category. The proportions found in one situation cannot be used for giving a probability in another situation where the constitution of the population is not known, as in most clinical situations.

The information has to be combined with data from other tests, and with qualitative indications on the Wechsler itself, for ex-

ample, differences on the Similarities (Hall, 1952) and "rotation" on the Block Designs (Shapiro, 1951), which are known to be related to the distinction being made. Another approach might be to attempt to eliminate individuals who come at the extreme of the distribution of DRs without having deteriorated. According to Wechsler (1944, p. 66), a DR of +20 is 2 Probable Errors from the mean, so that in a normal group one would expect to find about 9% above +20, and about 2% above +30. Subjects with the classical type of congenital reading disability, in which ordering and spatial organization are involved, might be expected to make poor scores on don't hold tests, especially Digit Span, Arithmetic, and Block Designs. There are three subjects in the present series with reading difficulties marked enough for it to be mentioned as one of their problems. They are all in the under-50 nonorganic group, and all have DRs greater than +29. It may be noted that the present group agrees with Wechsler's in the distribution at the other extreme of the scale. There are 8 minus ratios greater than 30 (3%) and 21 greater than 20 (8%).

Another thing associated with high DRs is low scoring on the test. Hold scores of 4 and don't hold scores of 2 do not suggest much absolute difference in the abilities concerned, but they would give a DR of +50. In the nonorganic under-50 group, there are 19 patients with IQ less than 80; 6 of these have DRs over +30, as opposed to 8 out of the other 129 with IQ over 79. This is not of much practical help, because low scores also occur in organic cases. In the organic under-50 group, there are 5 out of 23 with IQ below 80, and all 5 have DRs over +30. However, it does increase one's confidence in the significance of a high DR in a person of higher IQ.

As far as the under-50 nonorganic group is concerned, the study confirms the finding of Rogers (1950) that psychiatric groups show a higher DR than normals but do not differ among themselves. The over-50 nonorganic group, however, does not show more deterioration than the normal expectation. This curious age difference was also found, in psychiatric patients, by Garfield and Fey (1948). A similar age difference, in the same direction, appears in the organic groups. One explanation could be that the don't hold tests are affected by either mental illness or aging, but the two factors do not have an additive effect when they occur together, so that the older patient is not affected more than is allowed for in the age corrections; alternatively, the illness, when combined with greater age, might affect the hold as well as don't hold tests. If patients were retested after recovery with the equivalent form of the test, on each hypothesis the younger ones would be the same on the hold tests, and better on the don't hold; on the first hypothesis the older patients would be the same on both types of test, on the second, they would be better on both.

The fact that mental illness has this tendency to raise the DR does not destroy the utility of the index in the diagnosis of brain damage, since the latter raises it significantly more. An index of over 30 gives a strong initial suggestion of brain damage, especially when the IQ is not low. Whatever method of assessing deterioration is used, an estimate of intelligence is needed in evaluating the results, so that Wechsler's Scale can be used to fulfil a double purpose.

Summary

A modified form of Wechsler's Deterioration Ratio was calculated retrospectively for 261 male mental hospital inpatients who had been tested routinely over a period of 6 years. The scores were compared with the eventual diagnoses, and it was found that the ratio distinguished organic and nonorganic cases fairly well. In the nonorganic group, there was no variation corresponding to diagnosis, but there was confirmation of previous findings that psychiatric patients give larger ratios than normals. This, however, applied only to the younger group (under-50).

REFERENCES

Garfield, S. L., & Fey, W. F. A comparison of the Wechsler-Bellevue and Shipley-Hartford scales as measures of mental impairment. *J. consult. Psychol.*, 1948, **12**, 259–264.

Gutman, Brigette. The application of the Wechsler-Bellevue scale in the diagnosis of organic brain disorders. *J. clin. Psychol.*, 1950, **6**, 195–198.

HALL, K. R. L. Conceptual impairment in depressive and organic patients of the pre-senile age group. *J. ment. Sci.*, 1952, **98**, 256–264.

MCFIE, J., & PIERCY, M. F. Intellectual impairment with localized cerebral lesions. *Brain*, 1952, **75**, 292–311.

ROGERS, L. S. A comparative evaluation of the Wechsler-Bellevue mental deterioration index for various adult groups. *J. clin. Psychol.*, 1950, **6**, 199–202.

SHAPIRO, M. B. Experimental studies of a perceptual anomaly: I. Initial experiments. *J. ment. Sci.*, 1951, **97**, 90–110.

WECHSLER, D. *The measurement of adult intelligence.* (3rd ed.) Baltimore: Williams & Wilkins, 1944.

YATES, A. J. The validity of some psychological tests of brain damage. *Psychol. Bull.*, 1954, **51**, 359–379.

YATES, A. J. The use of vocabulary in the measurement of intellectual deterioration: A review. *J. ment. Sci.*, 1956, **102**, 409–440.

(Received April 29, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 239–244

VOCABULARY DEFICIT IN BRAIN OPERATED SCHIZOPHRENICS

ROY M. HAMLIN [1]
University of Pittsburgh School of Medicine
AND ELAINE F. KINDER
Rockland State Hospital, Orangeburg, New York

Traditionally, vocabulary has been regarded as one of the functions affected least and latest by impairment due either to schizophrenia or to brain damage. In a recent study, Smith and Kinder (1959; Smith, 1958) report no significant loss in vocabulary for topectomized schizophrenics as compared with controls, 8 years after surgery. On a variety of other tests the brain operated subjects did show significant losses. Smith has suggested that significant losses tend to appear more frequently on measures involving sustained attention and perceptual organization than on measures involving vocabulary and verbal skills.

Williams, Lubin, and Gieseking (1959) report seemingly opposite results for nonschizophrenic subjects with a variety of brain conditions: vocabulary and verbal skills not only show clearly significant losses, but furthermore the amount of loss is as much, or even more, than on spatial tests. Subject samples, site and nature of tissue damage, lapse of time following the lesions, and the type of vocabulary test used raise many questions. The present study focuses on the last of these three factors: what is the effect of an *oral* test format, as compared with a *multiple-choice* format, on vocabulary scores following brain surgery?

[1] Now at the Veterans Administration Hospital, Danville, Illinois. The senior author's participation in this research was made possible by support from Western Psychiatric Institute, Henry W. Brosin, Director; as well as by support from the Topectomy Follow-Up Project with a research grant (M-1191) from the National Institute of Mental Health and funds and assistance from the Research Foundation for Mental Health, Inc. of the New York State Department of Mental Hygiene.

This factor of test format, and of the testing procedure, has received little attention in studies of deficit following brain damage. The Williams et al. study does comment on the test used. These authors base their conclusions on the Reading and Vocabulary (RV) Test from the Army Classification Battery (ACB). They offer evidence for the essential equivalence of the RV test and the *Wechsler* oral Vocabulary test, citing correlations of .76 between these two tests for Army recruits, for enlisted men, and for brain injured patients. They also note that the "majority of the items in RV require the *S* to read and understand a paragraph. Only a minority are standard vocabulary items."

Although the present study limits consideration to test procedure, some comment on the general and often contradictory evidence for vocabulary deficit following brain damage should be offered. Two of the most relevant and adequate studies are those of Yates (1956) and of Weinstein and Teuber (1957). Yates, on the basis of an extensive review, concludes that "vocabulary does decline in patients suffering from brain-damage." This general conclusion may be supplemented by Weinstein and Teuber's careful consideration of differential effects when lesions involve specific brain areas. With pretest and posttest scores for both injured and control subjects, these latter authors report obvious loss in verbal and vocabulary scores some 10 years after focal lesions involving the left parietal-temporal areas, and/or with aphasia as a symptom. For other focal lesions, including frontal lobe lesions, they report little or no loss in vocabulary and similar tests.

In regard to the more immediate question of test format and test procedure, vocabulary deficit may well be in part a function of the vocabulary test used. In the previously cited study, Weinstein and Teuber (1957) found little verbal loss in certain brain injured groups; but these same subjects showed obvious loss on a hidden picture test. Is the multiple-choice test something like a hidden picture test, with the correct answer word "perceptually hidden" to some degree among three or four other words? Again, in the Smith and Kinder (1959) study, the Capps Homograph Test showed significant loss although oral vocabulary did not. The Homograph Test is a vocabulary test which introduces an additional element, requiring the subject to make a conceptual shift from one definition to another for the same word.

The present study used three test formats with subjects from the same sample studied by Smith and Kinder (1959): one oral, and two multiple-choice, tests of vocabulary. The first multiple-choice procedure *maximized* the factor of sustained attention. The second multiple-choice procedure *minimized* the factor of attention. The element of perceptual organization was assumed to be inherent in both multiple-choice procedures.

Since the subjects had taken Wechsler oral Vocabulary tests before surgery and again 8 years after surgery, the chief interest in the present study was in the new multiple-choice procedure which was designed to require both sustained attention and perceptual organization. The same words used in this multiple-choice procedure were then given orally and again under modified multiple-choice conditions to determine the comparability of these words to the Wechsler words, and to explore the extent to which each subject could achieve an approximate definition under any conditions. Three hypotheses were considered:

1. Topectomized schizophrenics will show significant loss on a multiple-choice test of vocabulary which *maximizes* the importance of sustained attention.

2. Topectomized schizophrenics will show no significant loss when the same words are then given as an oral vocabulary test. Smith and Kinder (1959) had already reported no loss in oral vocabulary for these subjects 8 years after surgery. The present oral test was designed to insure that the specific word list used did not account for a difference in results.

3. Topectomized schizophrenics will show a tendency to loss when the same words are given a third time in a multiple-choice procedure *minimizing* the importance of attention. The number of correct definitions was expected to be greater than on the first multiple-choice procedure and less than on the oral procedure.

Method

Subjects. The subjects and surgical procedures have been described in detail elsewhere (Lewis, Landis, & King, 1956; Smith & Kinder, 1959). The 40 subjects included in the present study were all patients at Rockland State Hospital, Orangeburg, New York. Of these 21 were operated subjects, and 19 nonoperated controls. All were from the New York State Brain Research Project, had been diagnosed as schizophrenic, and had been included in the study reported by Smith and Kinder.

The operations had been performed 10 years prior to the present study. The subjects had been originally divided into a C group of older subjects (age range 47–58 as of 1949) and a D group of younger subjects (age range 21–38 in 1949). Surgery consisted of an orbital or a superior topectomy, both bilateral frontal lobe operations.

Thirty-eight subjects were assigned to 19 pairs of operated and control patients, matched for preoperative Wechsler Vocabulary. In 15 pairs, it was possible to match the subjects by chronological age group. In the other 4 pairs, they were matched for Vocabulary but not for age.

Tests. In addition to the Wechsler Vocabulary before and 8 years after surgery, the 40 subjects were given three new vocabulary tests 10 years after surgery. The five tests, with the designation used in Table 1, were as follows:

1. Pre: The preoperative vocabulary score is the average of two Wechsler oral Vocabulary raw scores obtained within a period of a few months before surgery in 1949. On occasion, only one Wechsler was given before surgery, rather than the usual two.

2. Post-8: The average of two Wechsler oral Vocabulary tests given within a 30-day interval 8 years after surgery.

3. Oral: The 60 words in the first civilian edition of the Army General Classification Test (AGCT, 1947) were given orally and scored as in the Wechsler.

4. MC-1: The first multiple-choice procedure consisted of the same 60 items from the AGCT. The stem phrase was presented on a card for 5 seconds: for example, "BIG means the same as . . . ."

The stem card was then removed, and a 10-second delay followed. The four answer words from the AGCT were then presented, without the stem card, and the subject marked his response. The delay procedure emphasized the necessity for *sustained attention.*

5. MC-2: The second multiple-choice procedure employed the same materials. The four answer words from the AGCT were presented first, and the subject read them aloud. Mistakes in reading were corrected. The stem card was then presented and read aloud by the examiner, with the answer words still in view. The subject then gave his answer orally. At every step, the examiner encouraged maximum attention to each element in the task. This procedure minimized the effect of variable attention.

In the Oral, MC-1, and MC-2 procedures, a "base" and "ceiling" were used where obviously indicated; subjects responded to an average of some 45 of the 60 items. A correction for chance successes was used. The order of the tests was: MC-1, Oral, MC-2, with MC-2 given in a second session. The advantages and disadvantages of a rotated order for the tests were considered. With the number of subjects available and the anticipated variability of psychotic behaviors, it was felt that exactly the same procedure should be used with all subjects, in order not to lose the precision in analysis offered by pair matching. As previously indicated, the oral and second multiple-choice procedures were designed as control observations. The unexpected results with the second multiple-choice procedure will be discussed later.

Analysis. Results were analyzed by the sign test, and by t tests for differences between correlated means. p values refer to a one-tailed test.[2]

Results

Table 1 presents the results for the five vocabulary tests in terms of the number of controls surpassing topectomized patients, for 19 pairs of subjects matched on preoperative Wechsler Vocabulary. Significant values for the sign test are indicated. The results will be presented under the following headings: MC-1, Oral, MC-2, Wechsler, and the three hypotheses.

MC-1. The first multiple-choice procedure incorporated both *attention* and *perceptual organization,* as well as word knowledge. It was expected that this procedure might differentiate the topectomized and control groups. It failed to do so. Neither the sign test (Table 1) nor the difference between means ($t = 1.41$) is significant. As with every vocabulary test given after surgery, the *consistent tendency* for the controls to surpass the operates is apparent. The mean superiority of the controls is 4.3 words.

TABLE 1

Number of Controls Surpassing Topectomized Patients on Five Vocabulary Scores, for Nineteen Pairs of Subjects Matched on Preoperative Wechsler Vocabulary

	Pre	Post-8	MC-1	Oral	MC-2
Control Higher	8	14*	12	13*	16**
Ties	1	0	2	1	0

Note.—Pre = Preoperative Wechsler (1949).
Post-8 = Wechsler (1957).
MC-1 = Multiple-choice, with delay (1959).
Oral = New oral vocabulary (1959).
MC-2 = Multiple-choice, no delay (1959).

*$p = .05$.
**$p = .01$.

Oral. Since Smith and Kinder (1959) reported no significant difference in oral vocabulary for these subjects, it was expected that the new Oral vocabulary test would not differentiate the topectomized and control groups. This expectation was further encouraged by the high correlation between the new Oral and Wechsler oral given 2 years earlier: for 41 subjects, this correlation was .89. For 19 pairs matched on preoperative vocabulary, 13 pairs show Oral scores higher for the control subject, 5 higher for the operate, and 1 tie. The sign test is significant at the .05 level, suggesting some loss in oral vocabulary due to brain surgery. The t for the difference between means, however, is 1.29 and not significant. The average superiority of the control over the operate is 2.3 words.

MC-2. The second multiple-choice procedure minimized the importance of sustained attention but retained the factor of perceptual organization. This vocabulary test format yielded a clearly significant difference associated with brain surgery. Of 19 controls, 16 surpassed the topectomized subject with whom they were paired. The sign test is significant at the .01 level. The t for the difference between means is 3.21, also significant at the .01 level. The mean superiority of the controls over the operates is 6.3 words—

[2] Ardie Lubin, and Elizabeth Engle of the Walter Reed Army Institute of Research have provided the authors with detailed statistical analysis of all the major comparisons reported. In all statistics reported, their results confirm the figures contained in this paper.

nearly three times as much as for the same words given orally.[3]

The serial order in which the tests were given could be important: MC-1, Oral, MC-2. The control subjects showed a consistent increase in mean score on each later procedure: 30.9, 35.6, and 38.1. The topectomized subjects obtained the following mean scores on the respective tests in order of presentation: 26.6, 33.3, and 31.8.

Wechsler. Smith and Kinder, using analysis of covariance, report no significant loss in Wechsler oral Vocabulary, 8 years after surgery for these subjects. Pair matching, however, indicates significant loss. For 19 pairs, the control surpasses the operate in 14 cases. The sign test is significant at the .05 level. The t for the difference between means is 2.46, also significant at the .05 level.

To confirm this evidence for loss in Wechsler vocabulary, the data on all 52 subjects used in the previous study were reanalyzed by pair matching. It was possible to match each of the 24 control subjects with a topectomized subject on preoperative Wechsler Vocabulary. Inspection indicates that the 4 remaining operated subjects obtained comparable scores. The sign test (with the control surpassing the operate in 18 pairs) and the difference between means ($t = 2.42$) are both significant at the .05 level. The sign test just misses significance at the .01 level ($p = .011$). The relative mean loss due to brain surgery is 2.6 words.

Hypotheses. Results for the three hypotheses follow:

1. The topectomized schizophrenics do not show a significant degree of relative loss on the multiple-choice vocabulary test requiring both sustained attention and perceptual organization. As with all the vocabulary tests given after surgery, the consistent tendency for the controls to surpass the operates is suggested.

2. The topectomized subjects do show significant loss when the same words are then given orally. The sign test for the new Oral vocabulary indicates a difference in favor of the controls at the .05 level of confidence. The superiority of the controls need not be associated with the serial order in which the tests were given in the present study, since results for the Wechsler oral Vocabulary, given 2 years earlier, are comparable. Both the sign test and difference between means is significant at the .05 level.

3. The multiple-choice test which minimized the necessity for sustained attention clearly differentiates the operates from the controls. This test was given last in all cases. The topectomized schizophrenics show more loss, and more significant loss, in vocabulary on this test than on any of the others. The differential effect of "test procedure" is indicated, but both test format and order of presentation may contribute to the result.

Discussion

The present study considered specifically the effect of test format on vocabulary scores after brain surgery. The results do suggest that test procedure may be an important factor. However, another factor in test procedure, the order of test presentation, may influence the results as much as the test format. This Discussion will consider briefly the general results of the study, and will then comment on the special question of test procedure.

The evidence for some loss in vocabulary in topectomized patients, as compared with controls, 10 years after surgery is not surprising. The vocabulary losses reported here are small in terms of the instruments' precision: the relative loss in Wechsler Vocabulary is some 2.5 words. The number of studies, employing comparable controls, with which these results can be compared is exceedingly small. Only a few reports present: control subjects acceptably matched with experimental subjects on the basis of tests before brain damage; comparable pretest and posttest scores, especially after the lapse of a number of years; and any reasonably adequate evidence for the location of the lesion. Perhaps the most nearly comparable study is

[3] The gain from Oral to MC-2, for the 19 pairs of subjects, was greater for the control 15 times, for the operate 3 times, with 1 tie. This manipulation of scores has questionable justification, but does suggest the effect of different test formats on the same words.

that of Weinstein and Teuber (1957) previously cited. These authors reported marked loss on verbal and vocabulary tests some 10 years after certain focal brain injuries, but little or no loss on such tests when only the frontal lobes were involved. The table which they present does indicate some tendency to *relative* loss for every injury group considered, whatever area of the brain was involved. That is, some injury groups gained over their preinjury scores, but the gain was never as great as the gain made by the control subjects. The present results are similar: slight, but in this case sigiificant, loss in vocabulary 8 and 10 years after topectomies involving frontal lobe areas. Since Weinstein and Teuber (1957) report on a group test, on nonpsychotic subjects, and on injuries rather than bilateral operations, their study is only roughly comparable to the present one. The striking loss in verbal skills reported by Williams et al. (1959) must be evaluated in light of the heterogeneous brain conditions involved and the relatively short length of time following recovery. The evidence suggests that vocabulary deficit some 10 years after limited frontal lobe lesions is not great, but may be statistically significant under conditions such as those of the present study.

The effect of a specific test procedure employed in measuring vocabulary deficit has been given only incidental attention in previous studies. In one sense, the present results are clear. These topectomized subjects tend to show some deficit in vocabulary on every vocabulary test given 8 or 10 years after surgery. The effects of surgery are most clearly apparent, however, on the multiple-choice procedure MC-2, which was given last in order of testing. The effect of test procedure is indicated, but either test format or order of test presentation could account for the results.

In the present study, the effect of format and of order may be closely related. This suggestion calls for a final word on the special problem of demonstrating effects due to surgery when both experimental and control subjects are markedly psychotic. For such a purpose, the most suitable test may be one which minimizes the capricious and unpredictable vagaries of schizophrenic behavior, and yet retains some element related to brain damage. The MC-2 procedure may have been most effective for this reason. Each element in the task was presented separately, and every effort was made to insure active attention. Reading errors were corrected, and the subject reviewed each correction. When the crucial answer was finally permitted, the schizophrenic patient's sporadic or "functional" inattention to reality may have been largely overcome. The resulting optimal capacity to define words and deal with perceptual organization could then constitute a measure which would demonstrate the effects of surgery. The importance of simplifying elements, insuring attention, and using repeated measures should be considered. Paradoxically, such precautions might result in a measure of some aspect of "attention" which would reflect the effects of brain surgery.

The most discriminative test, MC-2, was given *last* to all subjects, after they had had repeated exposure to the words involved. Of the two oral tests, the one with which the subjects had had previous experience discriminated best. Sheer's (cf. Lewis, Landis, & King, 1956) astute observations on "practice gains" are related to the importance of driving each element in the task home before expecting schizophrenic subjects to respond in a sufficiently comparable manner to permit measurement of differences. That is, *familiarity* with a test may serve to decrease unpredictable and irrelevant schizophrenic behaviors. Unless the idiosyncratic capriciousness of schizophrenic behavior is taken into account, relevant differences may be readily obscured.

Summary

The study considered the effect of test procedure on measures of vocabulary deficit 8 and 10 years after bilateral topectomy. Three vocabulary tests, each employing the same words, were given to 21 operates and 19 controls. The three tests were a multiple-choice test maximizing the necessity for sustained attention, an oral test, and a multiple-choice test minimizing the importance of attention. The results and conclusions follow:

1. The topectomized schizophrenics showed a consistent tendency to vocabulary deficit on all tests, either oral or multiple-choice. The loss was statistically significant in most cases, but slight: for example, a mean loss of some two and a half words in Wechsler Vocabulary.

2. Vocabulary deficit was greater, and most significant, on the multiple-choice procedure which minimized attention, and which was given after the other two tests. This procedure differentiated the operates and controls at the .01 level of confidence. The mean superiority of the controls over the operates was nearly three times as many words as for the same words given orally.

3. The results suggest that the test procedure used may be an important factor in studies of vocabulary deficit. Relevant considerations include both the test format and the order of presentation. Both format and order may contribute to the results of the present study.

REFERENCES

Army General Classification Test. (First civilian edition) Chicago: Science Research Associates, 1947.

Lewis, N. D. C., Landis, C., & King, H. E. *Studies in topectomy.* New York: Grune & Stratton, 1956.

Smith, A. Changes in psychological test performances of brain operated schizophrenics after an 8-year interval. Unpublished doctoral dissertation, Yeshiva University, 1958.

Smith, A., & Kinder, E. F. Changes in psychological test performances of brain-operated schizophrenics after 8 years. *Science,* 1959, **129**, 149–150.

Weinstein, S., & Teuber, H. L. Effects of penetrating brain injury on intelligence test scores. *Science,* 1957, **125**, 1036–1037.

Williams, H. L., Lubin, A., & Gieseking, C. F. Direct measurement of cognitive deficit in brain-injured patients. *J. consult. Psychol.,* 1959, **23**, 300–305.

Yates, A. J. The use of vocabulary in the measurement of intellectual deterioration: A review. *J. ment. Sci.,* 1956, **102**, 409–440.

(Received May 9, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 245–248

AN EVALUATION OF THE NORTHWESTERN INFANT INTELLIGENCE TEST, TEST B

BERNARD B. BRAEN

Onondaga County Child Guidance Center

The Northwestern Infant Intelligence Test, Test B (Gilliland, 1951) contains 40 items and is designed for use with infants between the ages of 13 and 36 weeks. Preliminary statistical work with the test (Gilliland, 1951, p. 16) suggests that it may be a more suitable instrument for the assessment of infant intelligence than other tests for this age period. This paper attempts to provide a systematic assessment of reliability and validity of the test as well as other quantitative and qualitative considerations.

METHOD

The subjects consisted of 100 adoptive or boarding home babies of both sexes between the ages of 13 and 36 weeks inclusive.[1] In the first phase of the study the Cattell Infant Intelligence Test (Cattell, 1947) was administered to the baby by a qualified examiner.[2] The second phase occurred 3 days later when the Northwestern Infant Intelligence Test, Test B was administered to the baby by a different examiner.[3] The third phase involved the readministration of the Cattell by the original examiner [4] when the baby was 18 months of age.[5]

A frequency distribution of age by week for the babies administered the Northwestern is shown in Table 1. The mean age was 22.56 weeks with a standard deviation of 6.71 weeks.

RESULTS

The mean IQ for the Northwestern was 91.42 with a standard deviation of 9.54. The mean IQ for the Cattell was 112.00 with a

TABLE 1

FREQUENCY DISTRIBUTION OF AGE BY WEEKS FOR NORTHWESTERN-TEST B SUBJECTS

($N = 100$)

Age in Weeks	*F*
13	8
14	6
15	4
16	6
17	5
18	4
19	5
20	9
21	2
22	3
23	4
24	6
25	4
26	1
27	9
28	2
29	2
30	4
31	3
32	2
33	3
34	3
35	5
36	0

[1] The Child and Family Service and the Department of Public Welfare, Children's Division, both of Syracuse, New York were the participating agencies in the study.

[2] Deep appreciation is extended to Marilyn Rothschild for her crucial part in this project.

[3] The author administered all of the Northwestern tests.

[4] After 45 retests the original examiner left the employ of the clinic. The author administered the remaining 19 tests.

[5] The decision to retest the babies at 18 months with the Cattell was determined by two factors: (*a*) The correlation coefficient between the Cattell at 18 months and the Stanford-Binet (Form L) at 3 years was .67 (Cattell, 1947, p. 49). This coefficient suggests that the Cattell at 18 months appears to be measuring the same factor or factors as the Stanford-Binet at 3 years. (*b*) The results of this study were to have definite practical implications regarding the adoption testing program at the Child Guidance Center. For this reason it was important to gather validating data with expediency but without sacrificing rigor.

TABLE 2

MEANS AND STANDARD DEVIATIONS FOR NORTHWESTERN AND CATTELL STANDARD SCORE IQS FOR FOUR AGE GROUPS AND TOTAL GROUP

Age in Weeks	*N*	Northwestern		Cattell		Mean Difference	*t*
		M	*SD*	*M*	*SD*		
13–16	24	99.50	6.34	103.17	12.19	−3.67	−1.69
17–21	25	97.60	9.52	93.96	9.45	3.64	2.18*
22–27	27	103.89	8.35	102.59	8.87	1.30	.83
28–35	24	100.38	11.94	104.88	10.43	−4.50	−1.78
13–36	100	100.42	9.54	100.00	10.19	.42	.67

* Significant at the .05 level.

standard deviation of 15.21. On the basis of these findings it appears that the unequal means and standard deviations obtained with these tests make any direct comparison of IQs impossible. Further, without some sophistication with the concept of variability there may be a tendency to misinterpret the meaning of the IQ obtained with either test. In order to make comparisons between the two sets of scores, all the IQs on the Northwestern and Cattell were converted to standard scores with a mean of 100 and a standard deviation of 10. Also the 100 babies were divided into four age groups with approximately an equal number of subjects in each age group.

The converted means and standard deviations for the four age groups and the total group appear in Table 2.

From Table 2 it appears that variability of IQ from one test to the other occurs in all four age periods but the *t* tests reveal that it is most pronounced at the 13–16, 17–21, and 28–35 week levels. The *t* for the 28–35 week old group is artificially elevated due to the low ceiling on the Northwestern. By the IQ calculation method described by Gilliland (1951, p. 14) an infant of 36 weeks can only achieve an IQ of 112 on the Northwestern if he passes all 40 items. An infant of 35 weeks can only achieve an IQ of 115 if he passes all 40 items, and so on. Such a state of affairs could serve to lower artificially the mean IQ on the Northwestern for the 28–35 week old group, which would then contribute to a deceptive difference between the means of the Northwestern and Cattell.

The means and standard deviations of the difference between the Northwestern and Cattell standard score IQs for the four age periods are reported in Table 3.

It appears from these data that the mean difference in standard scores for each group and total age period is no less than 6 and no more than 10 units. However, the standard deviations for each age period suggest that individual infants can vary from no difference in relative position on the two tests to a difference of as much as 25 standard score units from one test to the other. Since only 3 days intervened between the Northwestern and Cattell testings, these data indicate that at this age level, examiner, subject, and test reliability are difficult to achieve.

Reliability

The reliability of the Northwestern for the 100 babies 13–36 weeks of age was assessed through the odd-even method. The resulting coefficient corrected by the Spearman-Brown formula was .95 ± .01.

Table 4 shows the odd-even reliability coefficients for the four age periods.

Gilliland (1951, p. 16) reported a corrected odd-even coefficient of .80 computed from

TABLE 3

MEANS AND STANDARD DEVIATIONS OF THE DIFFERENCES BETWEEN NORTHWESTERN AND CATTELL STANDARD SCORE IQS FOR FOUR AGE PERIODS

Age in Weeks	*N*	Range	*M*	*SD*
13–16	24	1–25	8.88	6.70
17–21	25	0–22	7.00	5.59
22–27	27	0–22	6.26	5.49
28–35	24	0–22	9.38	6.44

TABLE 4

NORTHWESTERN ODD-EVEN RELIABILITY COEFFICIENTS FOR FOUR AGE PERIODS

Age in Weeks	*N*	*r*	*SE*
13–16	24	.64	±.09
17–21	25	.76	±.05
22–27	27	.88	±.02
28–35	24	.89	±.02

data obtained from 214 babies between 13 and 36 weeks of age from the Chicago area. Cattell (1947, p. 49) reports odd-even coefficients for her test for age periods comparable to the Northwestern age levels. At the very early age levels (13 weeks and 3 months) both the Northwestern and Cattell are less reliable than at the later levels, but in relation to the Cattell at the early age period the Northwestern appears the more reliable instrument.

In spite of the fact that the Northwestern permits a reliable representation of particular skills during the testing session, the obtained coefficients do not give information regarding consistency of performance over time. It would appear then that the reliability of the test for predictive purposes is limited.

The reliability coefficients for each of the four age groups may in themselves be unreliable because of the small number of subjects at each level. However, since the progression of coefficients followed closely those obtained with the Cattell it appeared that this factor was not significantly affecting the resulting coefficients.

Correlational Analysis

The correlation coefficients between the Northwestern, Test B and the Cattell for the four age groups and the total group are reported in Table 5.

These coefficients suggest that there is an element of communality in the tests and that this communal aspect is apparent for all ages. The magnitude of the correlation coefficients remains about the same for each age period even though the reliability of both tests increases with age. Such a finding suggests that the tests have more in common at the 13–21 week period than the 22–36 week period.

The correlation coefficient of .58 between the Northwestern and Cattell for the whole age range may not reflect realistically the relationship between performance on the two tests because of the low ceiling on the Northwestern already described. In order to evaluate the effects of this artifact, the correlation was redone with the 28–35 week old group eliminated. The resulting coefficient was .62 ± .05. It does not appear that the elimination of the oldest group markedly affects the extent of the relation between performance on the two tests.

Sex Differences

A critical ratio was computed between the means for 61 males and 39 females on the Northwestern. The resulting ratio of 1.55 indicates that the obtained differences between the means of the sexes could be accounted for by chance.

Validity

Validity was assessed by correlating performance of 64 subjects on both the Cattell and Northwestern at 13 to 36 weeks with their performance at 18 months of age on the Cattell. Table 6 shows the means and standard deviations for the Northwestern and Cattell IQs at 13 to 36 weeks and the Cattell IQs at 18 months.

The correlation coefficient between the Northwestern IQs at 13 to 36 weeks and the Cattell at 18 months was .38 ± .07. The coefficient between the Cattell at 13 to 36 weeks and the Cattell at 18 months was .39 ± .07.

Both coefficients are low enough to indicate

TABLE 5

CORRELATION COEFFICIENTS BETWEEN NORTHWESTERN AND CATTELL IQS FOR FOUR AGE GROUPS AND TOTAL GROUP

Age in Weeks	*N*	*r*	*SE*
13–16	24	.51	±.10
17–21	25	.63	±.08
22–27	27	.54	±.09
28–35	24	.42	±.11
13–36	100	.58	±.04

TABLE 6

MEANS AND STANDARD DEVIATIONS FOR NORTHWESTERN AND CATTELL IQS AT 13–36 WEEKS AND CATTELL IQS AT 18 MONTHS

Test	*N*	*M*	*SD*
Northwestern (13–36 weeks)	64	90.91	9.38
Cattell (13–36 weeks)	64	111.92	16.91
Cattell (18 months)	64	110.58	10.65

that little faith can be placed in predictive statements regarding intelligence for the 13 to 36 weeks age group. In fact only about 14% of the variance on the Cattell at 18 months can be accounted for by variation in the Northwestern and Cattell IQs at 13 to 36 weeks. At the 13 to 36 week period, factors such as rapid and varying growth rates, emphasis on sensorimotor skills, varying motivation, relatively subjective administration and scoring procedures, and examiner unreliability all seem pertinent sources of uncontrolled variance that serve to reduce the predictive power of both the Northwestern and Cattell.

SUMMARY

This study was designed to investigate the reliability, validity, and certain other features of the Northwestern Infant Intelligence Test, Test B. One hundred adoptive or boarding home babies between the ages of 13 and 36 weeks were tested with the Cattell. Three days later the Northwestern was administered. The Cattell was readministered to 64 of the 100 babies when they were 18 months of age.

The general findings were:

1. The IQs obtained on the Northwestern and Cattell were not directly comparable due to unequal means and standard deviations.
2. The Northwestern has a low ceiling at the upper age levels which prevents full expression of an infant's developmental skills between the ages of 30 and 36 weeks.
3. Odd-even reliability for the Northwestern and Cattell is similar for the 13 to 36 week age period but the reliability for the Northwestern is somewhat better than the Cattell at the 13–16 week period. The reliability of both tests improved with increasing age.
4. There appears to be a common factor in both tests at the 13 to 36 week period. This is especially true at the 13–21 week level—probably due to a preponderance of sensorimotor items on both tests for this age.
5. There was no significant sex difference found on the Northwestern.
6. The validity results indicate that prediction of a child's intelligence level at 18 months is a very risky procedure when based on performance on the Northwestern and Cattell at the 13–36 week period.

REFERENCES

CATTELL, PSYCHE. *The measurement of intelligence of infants and young children*. New York: Psychological Corp., 1947.

GILLILAND, A. R. *Northwestern Intelligence Tests: Test B, for infants 13 to 36 weeks old*. Boston: Houghton Mifflin, 1951.

(Received May 11, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 249–252

JUDGMENTS OF ADJUSTMENT FROM TAT STORIES AS A FUNCTION OF EXPERIMENTALLY ALTERED SETS

BERNARD LUBIN [1]
Indiana University Medical Center

Several recent investigations have demonstrated the susceptibility of the projective techniques to situational influences (Masling, 1960). The arousal of temporary motivational states (Mussen & Scodel, 1955), the relationship between the subject and the examiner (Bernstein, 1956), the subject's attitude toward the test (Feldman & Graley, 1954), and the manner in which the test is defined (Henry & Rotter, 1956), produce identifiable changes in projective test response. These studies usually end at the point of measuring changes in the protocols; the implication seems to be that the modified protocols would influence subsequent clinical judgment.

The clinical psychologist usually works in a setting in which practical decisions having great consequences for the individual and for the community must be made. Often, the referrer requests an opinion as to an individual's "adjustment." Since the question rarely specifies a criterion situation for the judgment and since the various objective measures of adjustment are fairly specific (Tindall, 1955), the judgment is frequently based on a global estimate of intellectual and/or "emotional" controls. Psychoanalytic formulations continue to influence much of clinical practice; thus emotional control often is taken to mean control of sexual and aggressive drives, and inferences about psychological status are based upon the degree and manner in which drive derivatives enter consciousness and behavior.

The purpose of this investigation is to inquire into some of the determinants of global judgments of adjustment. Specifically, the effect of TAT stories which have been elicited under different instructions are judged for level of adjustment by clinical psychologists. Stories elicited under "Facilitating" instructions (prestige suggestion which placed high value on spontaneity and individuality) have already been shown to contain a higher amount of sexual and aggressive expression than stories elicited under "Neutral" or "Inhibiting" (prestige suggestion which placed high value on constraint and conformity) instructions (Lubin, 1960). The expectation is, therefore, that stories from the Facilitating condition will be judged as lower in adjustment than stories from either the Neutral or the Inhibited condition.

Method

In a previous investigation (Lubin, 1960), a random sample of 60 male college freshmen was tested in a 3×2 covariance design. The five conditions were: two Card conditions (two TAT cards with high pull of sexual content and two TAT cards with high pull of aggressive content), and three levels of Instructional Set (Inhibiting, Neutral, and Facilitating). Set was produced by means of prestige suggestion: the Facilitated group was told that "normal," "well-adjusted" people tend to let their imagination go as it is stimulated by the cards; the Inhibited group was told that normal, well-adjusted people are the master of their imagination and emotions; and the Neutral group was given innocuous instructions.[2] Subjects were randomly assigned to conditions, 10 subjects to a condition. Analyses indicated that Set produced a significant effect on sexual and on aggressive expression, and that the interaction between Set and Cards produced a significant effect on aggressive expression.

[1] I wish to thank James Norton for assistance with problems of design and analysis.

[2] A more detailed description of the methodology including the complete instructions can be found in the report of the previous investigation (Lubin, 1960).

Stimuli

The two experimentally treated TAT stories produced by each of the 60 subjects were stapled together. The stories were verbatim transcriptions and contained a notation as to reaction time. Within each envelope, the 120 stories (60 pairs) were divided into two sets of 30 pairs each, 60 stories told to TAT Cards 2 and 10, and 60 stories told to Cards 8BM and 20. The paired stories within each set were thoroughly scrambled.

Judges

Ten clinical psychologists, seven of whom possessed the PhD in clinical psychology and three who had a master's degree and at least 2 years of additional clinical experience, participated as judges.

Procedure

In the instructions, the judges were told that the stories were produced by 60 male subjects between the ages of 18 and 23. They were informed as to the TAT cards which had elicited the stories, and they were requested to rank the four major cues which they had used in rating the stories. Each judge received an envelope whose contents are described above. In addition, each envelope contained the following seven-point scale together with detailed instructions for its use.[3]

1. Well-integrated, happy person, socially effective
2. Only mild problems in essentially well-functioning person
3. Particular problems of some difficulty but social effectiveness maintained
4. Discomfort from problems severe enough to require therapy but ability to carry on
5. Acute neurotic problems but reality contact tenuously preserved
6. Severe neurotic problems with disorganization
7. Severe disturbance bordering on psychotic or prepsychotic

The seven-point scale was adapted from one developed and used by Dymond (1954) in an investigation of the effects of psychotherapy. Repeat-scoring reliability of this scale was found to be .94.

The two stories of each subject were assigned one rating by each of the 10 judges. Analysis of variance was conducted on the 600 ratings of adjustment.

Results and Discussion

It is important to note that the validity or accuracy of the ratings was not a subject of investigation. The stories had been produced by a sample of 60 male freshmen chosen randomly from the total population of male freshmen of a large university. It can be assumed that these subjects represented a sample of relatively well-functioning individuals. In order to preclude a skewed distribution of ratings, the judges were not told that the subjects were college students. However, neither were they told that the subjects represented a malfunctioning group, such as hospital or clinic patients. Under these conditions, one would expect that the effects of the stories themselves would be maximized.

The scatter plot revealed that the 600 ratings made by the 10 judges were distributed in a normal fashion over the seven scale points.

Table 1 presents the summary of the analysis of variance of the ratings of adjustment. It can be seen that the Instructional Set under which the TAT stories were elicited originally produced a significant effect on the judgments of adjustment. Neither Cards nor the interaction between Set and Cards was found to produce a significant effect on judgments of adjustment. Those aspects of the judgments determined by the sexual or aggressive stimulus patterns of the cards seemed to produce an effect of equivalent magnitude.

It seems reasonable to conclude, therefore, that the subject's attitude toward the testing situation can influence not only the test protocols (Lubin, 1960) but also may produce an effect on resulting clinical judgment.

Further information concerning the effect of Instructional Set upon judgments of adjustment is presented in Table 2. Since the scale was constructed so that higher scores represent judgments of poorer adjustment, the order of the means indicates that highest level of adjustment was rated for stories from the Inhibiting condition, lowest for stories from the Facilitating condition, with stories from the Neutral condition occupying the intermediate position. The Tukey Studentized Range Test reveals that the difference between the means of the Inhibiting and Facilitating conditions is large enough to be accepted with confidence, but that we cannot safely say where the Neutral condition falls in between the other two.

[3] A copy of the complete instructions has been deposited with the American Documentation Institute. Order Document No. 6631 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

ANALYSIS OF VARIANCE OF RATINGS OF ADJUSTMENT

Source	*df*	*SS*	*MS*	*F*
Between Sets	2	688.53	344.27	4.561*
Between Cards	1	7.34	7.34	.097
Set × Cards	2	30.41	15.21	.201
Subjects Within Set × Card Cells	54	4076.30	75.49	
Totals	59	4802.58		

* Significant at the .05 level.

Although the lists of story cues upon which the judges based their ratings were too varied to permit meaningful categorization and analysis, 8 of the 10 judges mentioned "amount of sex or aggression" as one of the four main cues which they used. The previous investigation (Lubin, 1960) demonstrated that Instructional Set produced a significant effect on both sex and aggressive expression and further analysis of the effect indicated that the effect was linear, i.e., highest expression of sex and aggression in the Facilitating condition, lowest in the Inhibiting condition, and an intermediate position for the Neutral condition. Thus there is a strong suggestion that judgments of adjustment based on TAT stories are significantly influenced by the amount of sex and aggression which is expressed.

These findings support the observations of Soskin (1954) that when the clinical psychologist is requested to make interpretations based on projective test protocols, his judgments tend to be biased in the direction of pathology. Also, Kenny and Bijou (1953) found that when clinical psychologists were asked to rank TAT stories according to their interpretive significance, they tended to give greater weight to contents which represented expressions of the sexual and aggressive drives.

The findings of this investigation point to some of the risks involved in "blind analysis." When the subject has a Set to be spontaneous and to individualize himself, he expresses more sex and aggression in his TAT stories (Lubin, 1960), and when the clinical psychologist, without knowledge of the subject's Set, makes a judgment of the subject's adjustment based on these same TAT stories his judgment tends to be biased in the direction of pathology to the extent to which sex and aggression is expressed in the stories.

It might be objected that in practice the clinical psychologist does not make judgments of adjustment based solely on the TAT and certainly not on such a small number of cards as were used in this study. The first part of the objection must be granted; such an important judgment would be based upon a battery of tests rather than a single instrument. It should be noted, however, that within the test battery, the TAT is likely to be used in a variety of ways (Dana, 1956). In addition, interpretation of the TAT in the clinical setting is more often based on global judgment than on the few time consuming objective methods of analysis which have been proposed.

SUMMARY

In order to test the hypothesis that judgments of adjustment made by clinical psychologists are influenced by the Set under which a subject takes a projective test, two TAT stories from each of 60 subjects were rated on a seven-point scale of adjustment by 10 clinical psychologists. The stories previously had been elicited under differential Instructional Sets: Inhibiting ($N = 20$), Neutral ($N = 20$), and Facilitating ($N = 20$).

TABLE 2

MEAN RATINGS OF ADJUSTMENT BY SET CONDITION

Set	Mean
Inhibiting	35.35
Neutral	38.35
Facilitating	43.55

It was found that Instructional Set significantly influenced the ratings: most pathology was rated for stories from the Facilitating condition and least for stories from the Inhibiting condition. Collateral data suggested that the amount of sexual and aggressive expression in the stories was the most important factor which influenced the ratings.

REFERENCES

BERNSTEIN, L. The examiner as an inhibiting factor in clinical testing. *J. consult. Psychol.*, 1956, **20**, 287–290.

DANA, R. H. Selection of abbreviated TAT sets. *J. clin. Psychol.*, 1956, **12**, 36–40.

DYMOND, ROSALIND F. Adjustment changes over therapy from Thematic Apperception Test ratings. In C. R. Rogers & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954.

FELDMAN, M. J., & GRALEY, J. The effects of an experimental set to simulate abnormality on group Rorschach performance. *J. proj. Tech.*, 1954, **18**, 326–334.

HENRY, EDITH M., & ROTTER, J. B. Situational influences on Rorschach responses. *J. consult. Psychol.*, 1956, **20**, 457–462.

KENNY, D. T., & BIJOU, S. W. Ambiguity of pictures and extent of personality factors in fantasy responses. *J. consult. Psychol.*, 1953, **17**, 283–288.

LUBIN, B. Some effects of set and stimulus properties on TAT stories. *J. proj. Tech.*, 1960, **24**, 11–16.

MASLING, J. The influence of situational and interpersonal variables in projective testing. *Psychol. Bull.*, 1960, **57**, 65–85.

MUSSEN, P. H., & SCODEL, A. The effects of sexual stimulation under varying conditions on TAT sexual responsiveness. *J. consult. Psychol.*, 1955, **19**, 90.

SOSKIN, W. F. Bias in postdiction from projective tests. *J. abnorm. soc. Psychol.*, 1954, **49**, 69–74.

TINDALL, R. H. Relationships among indices of adjustment status. *Educ. psychol. Measmt.*, 1955, **15**, 152–162.

(Received May 12, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 253–256

ORTHOPEDIC DISABILITY AS A FACTOR IN HUMAN-FIGURE PERCEPTION

AURELIA LEVI
Teachers College, Columbia University

There is perennial interest in the possible relationship between personality dynamics and various features of human-figure drawing. Though experimental studies frequently fail to support such a relationship, the impression continues that it does nevertheless exist, and an occasional study showing positive results helps to keep the impression alive.

Some of this ambiguity of result may be owing to the fact that even if it be true that a figure-drawing is a projection of its creator's body image (Machover, 1949), the projective process is necessarily limited by perceptual limitations, including perceptual prejudices (Postman, Bruner, & McGinnies, 1948). Indeed, this is the heart of the basic projective hypothesis. It follows then that before we can proceed to evaluate the peculiarities of a drawing—its omissions, overemphases, distortions, or whatever—in terms of its creator's dynamics, we would do well to establish the prejudicial influence of these dynamics on his perceptions. In other words, even assuming that a drawing is a projection, before we can attribute a bit of elaborate overemphasis in the drawing to its creator's undue preoccupation with the part in question, we have to insert a middle step and show that his perception—as yet uncomplicated by the act of bodying forth mental images through paper and pencil—has already been shaped by that particular preoccupation. This point has been made concisely by Silverstein and Robinson (1956): "The assumption that the physical body, the body image, and the drawn figure are in isomorphic relation" remains as yet unjustified. From their study they conclude that "this one-to-one relationship does not seem to exist" (p. 340).

It is only with the existence of prejudicial influences on the perceptions of several diagnostic groups that the present study is concerned. The final step, relating these perceptions to the recreative process of figure-drawing, is beyond the scope of this study. Our task will be to separate Silverstein and Robinson's three isomorphs into two groups of two, and investigate a hypothesized one-to-one relation between the first two, the physical body and perceptions of the human figure (the body image). Our diagnostic groups differ from each other with respect to orthopedic disability—a comparatively simple, objective, distinguishing criterion.

This study hypothesizes a one-to-one relation between the physical disability and perceptions of the human figure, and that a particular orthopedic disability acts as a dislocating influence on perceptions of the relevant body part. Specifically, it is hypothesized that subjects with disability of the legs will be unusually sensitive to the legs in drawings of the human figure; and that subjects with disability of the arms will be unusually sensitive to the arms in drawings.

Since the group of back disabilities is a less homogeneous group than the other two, and has less well-defined etiologies which have been thought to be of psychic origin, it is further hypothesized that this group will show a greater resemblance to the characteristics of a control group of nondisabled.

PROCEDURE

Subjects

The experimental group was composed of 38 subjects with orthopedic disabilities as follows: (*a*) 12 with traumatic disability (fractures, amputations) to the arms, and no other disability; (*b*) 13 with traumatic disability to the legs, and no other disability; (*c*) 13 with a variety of low-back disabilities (spinal

TABLE 1

AGE, SEX, AND EDUCATION OF CONTROL AND DISABLED SUBJECTS

Subjects	*N*	Age		Sex		Education	
		$\bar{X}$	Range	M	F	$\bar{X}$	Range
Control	35	40.8	19–68	21	14	9.0	2–20
Disabled	38	41.9	16–67	28	10	8.5	0–20

fusion; laminectomy; arthritis of the lower back; uncomplicated, nonradiating low-back pain of undetermined origin), and no other disability. The control group was composed of 35 subjects without any orthopedic disability. As shown in Table 1, the age of the experimental group ranged between 16 and 67, mean age 41.9, and of the control group between 19 and 68, mean age 40.8. The number of years of education for the experimental group ranged from 0 to 20, mean 8.5 (median 9), and for the control group between 2 and 20, mean 9.0 (median 10).

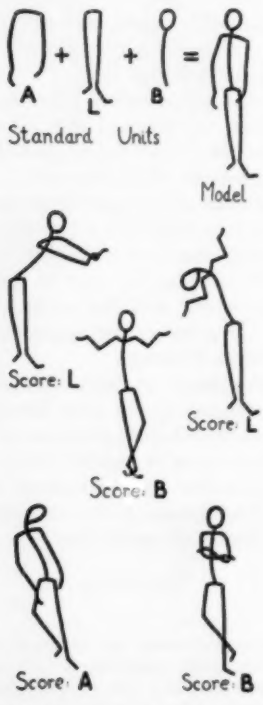

FIG. 1. Model figure, standard body units, and five examples of stick figures containing one standard unit each.

Instrument

The instrument was a set of 27 cards, each bearing a pair of stick figures (after Sarbin, 1954; Sarbin & Hardyck, 1955), and a single large card bearing a model stick figure. All the stick figures are composed of three parts: an arm unit, a leg unit, and a head-and-spine unit. The three parts making up the model are considered to be the standard units of reference; each of the 54 paired figures contains only one standard unit, the other two units being variants (Figure 1). Each stick figure appears twice in the set, once on the left and once on the right. The figures are so paired that 18 of them contain the standard arm unit, 18 contain the standard leg unit, and 18 contain the standard head-and-spine unit, but no 2 figures on one card contain any identical unit. Thus, from an objective point of view, each figure bears exactly the same amount of resemblance to the model as every other figure.

Method

Each subject was instructed as follows: "Here is a model figure, on the large card. I am going to show you other figures in pairs, and I want you to pick out which one of each pair looks more like the model, or resembles the model more closely." Scores for each subject are the number of choices he made involving each standard unit, the maximum number of choices possible for any unit being 18, and the minimum being 0. For example, a figure choice is credited to the Arm score if the standard unit it contains is the standard arm unit; similarly, it is credited to the Back score, if the standard unit it contains is the standard head-and-spine unit; etc. (Figure 1).

TABLE 2

COMPARISON OF MEANS OF THREE DISABILITY GROUPS WITH A CONTROL GROUP

Group	*N*	Mean	*t*[a]	*p*
A Choices				
Group A	12	14.17	3.49	.01
Control	35	8.94		
L Choices				
Group L	13	11.23	5.05	.001
Control	35	7.49		
B Choices				
Group B	13	12.92	1.48	.20
Control	35	10.54		

[a] The *t* tests comparing Groups A and B with the control group were computed by a method designed for groups whose variances are unknown but presumed equal (see Table 3); the *t* test comparing Group L with the control group was computed for the case for which variances are unknown but presumed unequal (Walker & Lev, 1953).

TABLE 3

COMPARISON OF VARIANCES OF THREE DISABILITY GROUPS WITH A CONTROL GROUP AND WITH EACH OTHER

Group	*N*	Variance	*F*	*p*
A Choices				
Group A	12	20.7	1.02	*ns*
Control	35	20.29		
L Choices				
Group L	13	3.19	3.35	.05
Control	35	10.67		
B Choices				
Group B	13	18.41	1.47	*ns*
Control	35	27.02		
Group A	12	20.7	6.49	.01
Group L	13	3.19		
Group B	13	18.41	5.77	.01
Group L	13	3.19		
Group A	13	20.7	1.12	*ns*
Group B	12	18.41		

RESULTS

As was hypothesized, the arm-disabled and leg-disabled groups show a perceptual sensitiveness that varies with the site of disability. For these two groups, means differ from those of nondisabled persons (Table 2), at a high degree of significance (.01 and .001, respectively).

As was hypothesized, the group of back disabilities shows no such perceptual difference from the nondisabled. The mean B Score for the back-disabled group does not differ significantly from the control mean (Table 2).

An unexpected finding is that the leg-disabled group is far more homogeneous than any of the others (Table 3).

DISCUSSION

The main finding of the study—that there is a one-to-one relationship between certain types of physical disability and perceptions of the human figure—provides a necessary, though not sufficient, step for a demonstration that a figure-drawing is a projection of its creator's body image.

That this one-to-one relationship is not a simple, across-the-board condition whose existence can be assumed for every kind of physical idiosyncrasy may be seen from the fact that the group of back disabilities, whose origins and symptomatology are much less clearcut than those of the other two groups, shows a less distinct perceptual prejudice; and from the differing variability of the groups.

It may be that the degree to which the disability is visually apparent immediately, exerts an important influence. Under such a scheme, low-back sufferers would be at the low end of a continuum of immediately apparent crippling; arm fractures and even amputations of the hand or arm—especially when fitted with a cosmetic prosthesis—while becoming apparent enough in any situation involving interaction, would nevertheless be more readily concealable than the impaired and distorted ambulation of the leg-disability group. Such an explanation might account for the findings with respect to the variability of the groups as well as their perceptual prejudice.

It should be noted that the results of this study may possibly be dependent on the fact that the control group and all three experimental groups were fairly homogeneous with respect to age, education, and socioeconomic level. To the extent to which differences in these factors might exert competing influences—as, for instance, the possibility that an extremely gifted, high *n* Achievement subject might bring into play an opposing perceptual prejudice akin to repression (Postman, Bruner, & McGinnies, 1948)—the pattern of results might be expected to vary; such an effect would, of course, be a matter for clarification through further research.

SUMMARY

This study compared three physical-disability groups with a group of nondisabled controls for their perceptual reactions to a structured test involving resemblances among schematized human-figure drawings.

It was ascertained that subjects with arm disability are particularly sensitive to the arms in a drawing, and that subjects with leg

disability are particularly sensitive to the legs in a drawing. Subjects with a variety of low-back ailments—often thought to be psychogenic—appear to be closer to the nondisabled control in their reactions than do either of the other two groups.

The group of leg disabilities is a much more homogeneous group than either the other two disabled groups or the controls.

A possible explanation that would account for all these findings is offered on the basis of the degree to which the disability is instantly apparent.

It is considered that this study gives support to a hypothesized one-to-one relationship between the physical body and the body image.

REFERENCES

MACHOVER, KAREN. *Personality projection in the drawing of the human figure.* Springfield, Ill., Charles C Thomas, 1949.

POSTMAN, L., BRUNER, G., & MCGINNIES, E. Personal values as selective factors in perception. *J. abnorm. soc. Psychol.*, 1948, **43**, 142–154.

SARBIN, T. R. Role theory. In G. Lindzey (Ed.), *Handbook of social psychology.* Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 223–258.

SARBIN, T. R., & HARDYCK, C. D. Conformance in role perception as a personality variable. *J. consult. Psychol.*, 1955, **19**, 109–111.

SILVERSTEIN, A. B., & ROBINSON, H. A. The representation of orthopedic disability in children's figure drawings. *J. consult. Psychol.*, 1956, **20**, 333–341.

WALKER, HELEN M., & LEV, J. *Statistical inference.* New York: Holt, 1953.

(Received May 18, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 257–259

TAT PERFORMANCE AS A FUNCTION OF ANXIETY AND COPING–AVOIDING BEHAVIOR[1]

E. JERRY PHARES
Kansas State University

A characteristic often associated with anxiety is its potential for generalization. Based perhaps on either qualitative similarity or identical elements, anxiety may become attached not only to the original arousal situation but also to other situations.

One such situation is that of projective testing. Anxious patients frequently seem to project threat into the test stimuli. For example, Rotter (1940, 1946) has stated that anxiety on the TAT is revealed in plots that emphasize sudden physical accidents and emotional trauma. Similarly, Schwartz (1955), from a Freudian viewpoint, has related the expression of castration anxiety on the TAT to themes in which occur genital or other body injury or loss, sexual or personal inadequacy, intrapsychic or extrapsychic threat, and loss of cathected objects.

However, to expect every anxious patient to produce such themes would seem a too simple approach to the problem. Not all anxious people handle their anxiety alike. Some subjects, having perceived a threatening stimulus, become evasive and produce bland stories, while others respond directly and create themes indicative of threat. The distinction drawn here parallels that of the perceptual defense-sensitization dimension (Carpenter, Weiner, & Carpenter, 1956).

In a similar vein, Weisskopf-Joelson, Asher, and Albrecht (1957) investigated "label-avoidance" as a manifestation of repression. They found some support for the hypothesis that people who strongly repress an impulse tend to avoid expressing this impulse in situations carrying its label to a high degree. Applied to the TAT this could mean that subjects with strongly repressed sexual or aggressive impulses would frequently fail to give sexual or aggressive themes to pictures suggestive of an aggressive or sexual content.

The hypothesis of the present study is that high anxious subjects will show greater preference for TAT themes involving accident, threat, or trauma than will low anxious subjects when matched for tendency to evade or cope with threatening stimuli. The present study resembles one by Goodstein (1954) who found a nonsignificant relationship between anxiety and preference for anxiety-like TAT statements without controlling for coping-avoiding tendencies.

METHOD

Selection of High and Low Anxious Subjects

The Taylor anxiety scale (*A* scale) (1953) was used to select high and low anxious subjects. From 263 general psychology students who took a 50 item form of the *A* scale during pre-enrollment, 25 high anxious (scores above 18) and 25 low anxious females (scores under 7) were selected.

Determination of Coping-Avoiding Tendencies

This technique stems from a distinction between copers and avoiders recently made by Mainord (1956). He demonstrated that copers recalled more nonsense syllables associated with disturbing words than with neutral words, and avoiders recalled more syllables associated with neutral words. Goldstein (1959) utilized the same distinction in predicting differential responses to fear-arousing propaganda. Both of these investigations used an incomplete sentences blank (ISB) consisting of 40 critical and 20 filler stems. The former have direct sexual and aggressive implications. Critical items are scored on a three-point scale in terms of specificity, strength, and arbitrariness of response. A subject's score is the sum of

[1] Based on a paper delivered at the Annual Meeting of the Midwestern Psychological Association, St. Louis, Missouri, April 29, 1960.

This study was supported in part by a grant from the Faculty Research Fund of Kansas State University.

the weights assigned to each item and a high score indicates coping tendencies. Copers are thus subjects who respond most directly to the implications of the stems while avoiders are those who respond most evasively. Using this technique, two judges independently scored 13 protocols from a pretest population with 89% scoring agreement.

TAT Measure

To increase objectivity, the usual method of administering and scoring the TAT was modified. Seven cards (4, 6BM, 7BM, 13MF, 14, 17BM, and 20) were presented, each accompanied by six themes: four neutral and two embodying threat, accident, or trauma. For example, with 6BM:

neutral—This young man has come to the mansion to apply for the job of gardener. He has been asked to wait by this elderly maid. So, hat in hand, he waits.

threat—The son has just had to tell his mother that all of their savings have been wiped out. The stocks are worthless that they invested in. They are both stunned by this development. Nothing is left.

Each subject rank ordered the alternative themes in terms of how well they fitted the card. The subject's score is the sum of the ranks of the 14 threat themes. The higher the score, the lower the preference ranking for threat themes.

Each subject was tested individually. The modified TAT was administered, followed by the coping-avoiding ISB. Subjects responded anonymously to both TAT and ISB.

Results and Discussion

From the foregoing procedure it was possible to match high and low anxiety subjects with respect to ISB scores. The N of each group was 19 and the mean ISB score for each group was 35.0.

The mean TAT score for the high anxious group was 47.2 (SD 4.7), and for the low anxious group 53.6 (SD 6.5). Applying a t test [2] the difference between the means is significant and in the predicted direction ($t = 3.4$, $p = .001$).

The results bear out the hypothesis that anxious people see more threat in TAT cards than do nonanxious people when their technique for defending against anxiety is controlled. Goodstein (1954) reported a nonsignificant trend in a similar direction due perhaps to a lack of control for the coping-avoiding dimension. Although a portion of his data was based on 10 anxious and 10 nonanxious subjects while the present data is based on 19 pairs of subjects, the discrepancy between probability estimates in the two studies is much greater than would be expected merely by augmenting N.

Thus, it seems probable that in a given unselected population of anxious subjects, some who are copers and others avoiders, the confirmation of an otherwise perfectly logical and tenable hypothesis might be prevented. For example, Lesser (1959) demonstrated that under conditions of high anxiety about aggression, the intercorrelations among various measures of aggression were significantly lower than in the case of low anxiety over aggression. A behavior occurs not simply as a function of one variable but as a function of the relationship among several variables.

Generalization of the present results is, of course, limited by the sex composition of the population and by the modified TAT procedure.

[2] A one-tailed test of the distribution of t was used.

Summary

This study hypothesized that anxious subjects will show greater preference for TAT themes involving accident, threat, or trauma than will nonanxious subjects when matched for tendency to evade or cope with threatening stimuli.

Twenty-five subjects high on the Taylor anxiety scale and 25 low on that scale were administered a modified TAT and a special ISB which measures coping-avoiding tendencies. Subjects rank ordered neutral and threatening themes which accompanied seven TAT cards in terms of how well they fitted the cards.

With this procedure, 19 pairs of high and low anxiety female subjects matched for coping scores confirmed the hypothesis at a statistically significant level.

REFERENCES

Carpenter, B., Weiner, M., & Carpenter, Janeth. Predictability of perceptual defense behavior. *J. abnorm. soc. Psychol.*, 1956, **52**, 380–383.

Goldstein, M. J. The relationship between coping and avoiding behavior and response to fear-arousing propaganda. *J. abnorm. soc. Psychol.*, 1959, **58**, 247–252.

GOODSTEIN, L. D. Interrelationships among several measures of anxiety and hostility. *J. consult. Psychol.*, 1954, **18**, 35–39.

LESSER, G. S. Population differences in construct validity. *J. consult. Psychol.*, 1959, **23**, 60–65.

MAINORD, W. A. Experimental repression related to coping and avoidance behavior in the recall and relearning of nonsense syllables. Unpublished doctoral dissertation, University of Washington, 1956.

ROTTER, J. B. Studies in the use and validity of the Thematic Apperception Test with mentally disordered patients: I. Method of analysis and clinical problems. *Charact. Pers.*, 1940, **9**, 18–34.

ROTTER, J. B. Thematic Apperception Tests: Suggestions for administration and interpretation. *J. Pers.*, 1946, **15**, 70–92.

SCHWARTZ, B. J. The measurement of castration anxiety and anxiety over loss of love. *J. Pers.*, 1955, **24**, 204–219.

TAYLOR, JANET A. A personality test of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48**, 285–290.

WEISSKOPF-JOELSON, EDITH, ASHER, E. J., & ALBRECHT, K. J. An experimental investigation of "label-avoidance" as a manifestation of repression. *J. proj. Tech.*, 1957, **21**, 88–93.

(Received May 18, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 260-266

RELATIONSHIPS BETWEEN DESCRIPTIVE CONTENT AND INTERACTION BEHAVIOR IN INTERVIEWS[1]

JEANNE S. PHILLIPS, RUTH G. MATARAZZO, JOSEPH D. MATARAZZO, GEORGE SASLOW
University of Oregon Medical School
AND FREDERICK H. KANFER
Purdue University

This study was undertaken in order to explore possible relationships between two different general approaches to the description and measurement of verbal interview behavior. One widely applied frame of reference directs attention to the communication aspects of verbal behavior, that is, to some symbolic dimension of the content of the words spoken, using content analysis to define and quantify its variables. A second frame of reference focuses upon quantitative *temporal* characteristics of interview interactions, utilizing measures such as number and duration of utterances, duration of silences, etc.

The few studies which have incorporated measures of *both* content *and* temporally defined variables have usually indicated the greater discriminatory power of the latter. For example, Page (1953), Lennard (1955), Goldman-Eisler (1952), and others have found that quantity or tempo of verbal output is both more stable and more highly correlated with various criteria of personal adjustment or psychotherapeutic success than are content-derived variables.

Lundy (1950) found that a single patient seen concurrently by two therapists differing in therapeutic technique showed no difference in his responses to the two therapists in measures of content (Distress-Relief Quotient, Raimy's self-references, etc.). However, differences in *tempo* of interaction in the two sets of therapeutic interviews were apparent in the protocols. As a secondary aspect of a subsequent study, Lundy (1955) compared the temporal variables with clinical estimates of occurrence of significant emotional and topical content in "key" interviews. The results indicated that the contentual and temporal variables were related.

The present study involves the direct investigation of relationships between more precisely measured content variables and the temporal measures of the Interaction Chronograph. This instrument and Chapple's underlying interaction theory (Chapple, 1949; Chapple & Arensberg, 1940) constitute an extensive, systematically developed attempt to describe temporal phenomena in verbal behavior and to investigate their significance. Definitions of the Interaction Chronograph variables involved in the present study are given in Table 1. Without more extensive knowledge on our part of the subtle differences in "meaning" of the several intercorrelated Interaction Chronograph variables (Matarazzo, Saslow, & Hare, 1958), precise hypotheses as to specifically which content and interactional variables would be related could not be formulated. However, tentative hypotheses regarding the general types of relationships to be expected on a "face validity" basis were obviously the determinants of the kinds of content categories selected for use,

[1] This investigation was supported by a research grant (M-735) from the National Institute of Mental Health, of the National Institutes of Health, United States Public Health Service. This paper is based in part on a dissertation, under the direction of Frederick H. Kanfer, submitted by the senior author in partial fulfillment of the requirements for the degree of doctor of philosophy at Washington University. The data were collected while all of the investigators were at Washington University.

TABLE 1

INTERACTION CHRONOGRAPH VARIABLES

Pt.'s Units	The number of times the patient acted
Pt.'s Tempo	The average duration of each action plus its following inaction (silence), as a single measure
Pt.'s Silence	The average duration of the patient's silences
Pt.'s Adjustment	The durations of the patient's interruptions minus the durations of his latencies in responding, divided by Pt.'s Units
Pt.'s Initiative	The percentage of times, out of the available number of opportunities (usually 12) in Period 2, in which the patient acted again (within a 15-second limit) following his own last action
Pt.'s Quickness	The average length of time in Period 2 that the patient waited before taking the initiative following his own last action
Pt.'s Dominance	The number of times in Period 4 that the patient "talked down" the interviewer minus the number of times the interviewer talked down the patient, divided by the number of Pts.'s Units in that period

and are illustrated below in the description of the content system.

PROCEDURE

Forty patients randomly selected from new referrals to the Psychiatric Outpatient Clinic of a large urban medical center were interviewed by the same psychiatrist according to the published rules of the partially standardized interview which is used with the Interaction Chronograph in order that the interviewer may serve as a partially controlled or independent variable during each of five predefined subperiods of the interview (Matarazzo, Saslow, & Matarazzo, 1956; Saslow & Matarazzo, 1959). Each interview was observed from an adjoining room through a one-way mirror by the same observer, who recorded the ongoing interaction on an Interaction Chronograph and made a verbatim sound recording on a Gray Audograph. Ten of the 40 patients were omitted from the final experimental sample because of deficiencies in the sound recordings, or, in a few cases, because of the presence of acute psychosis which resulted in confused, incoherent interview content. The 30 remaining patients served as the subjects of the present study. Of these, 17 were female, 13 were male. Their ages ranged from 20 to 61 years, with a median age of 35.5. The most frequent diagnoses were: hysteria (eight cases), anxiety neurosis (seven cases), depression (six cases) and schizoid personality (four cases).

The sound recordings were carefully and repeatedly monitored by the typist and a judge, until a high degree of recording transcription fidelity was achieved. Since the reliability of the content scoring process had been demonstrated on an independent sample of transcripts shortly before (Phillips, 1957), the final transcripts were unitized and then categorized, unit by unit, by one experimenter.

The content aspects of the verbal interview behavior were defined and quantified according to the category system schematically diagrammed in Table 2. The system and its development have been described in Phillips (1957). It represents an adaptation and extension of the Interpersonal System devised by Freedman, Leary, Ossorio, and Coffey (1951), C and C′ in Table 2, which consists of a circular continuum of 16 categories, representing qualitative blendings of two orthogonal dimensions, love-hate and dominance-submission. A seventeenth category was added for coding of units unscorable or neutral within the Interpersonal System.

In addition to the Interpersonal System, several other dimensions were coded in order to achieve completeness of coverage and more general applicability. These other dimensions were operationally defined by: (*a*) coding of units without interpersonal reference, D and D′ in Table 2, such as "I sat down and ate," "My head aches," etc.; (*b*) coding of the actor or subject of a description and of the receiver of action or attitude, if any, A and E in Table 2; and (*c*) coding of the general topic discussed, e.g., marriage, finances, symptoms, etc. Provision was also made for differentiation of descriptions of *motor acts*, C and D, from description of *nonmotor states* of being, thinking, feeling, etc., C′ and D′. For example, "I yelled at her" is an interpersonal act, C, while "I was angry at her" is coded as an interpersonal state, C′; "I ate dinner" is a noninterpersonal act, D, while "I felt sick" is a noninterpersonal state, D′. Those units which referred directly to the ongoing interview interaction and had no referent outside of the current situation (e.g., "Thank you, doctor," "I agree with you") were classified separately, according to their function, within categories adapted from Bales (1950), II in Table 2. A residual category was utilized for unscorable units, III.

The resulting system is of a pyramidal nature, involving several series of mutually exclusive categories. It was considered particularly suited to the purposes of this study because of its interpersonal interactional emphasis and its use of content variables seemingly parallel to the temporal variables of the Interaction Chronograph. For example, it was expected that one or more of the Interaction Chronograph measures of verbal "output" (number and duration of utterances, etc.) would be related to the frequency of content describing the self as physically *active*, Self C + D in Table 2. Since the former measures of verbal output in the interview have been

TABLE 2

CONTENT CATEGORY SYSTEM

I. Description (units which "tell about" an event)
- A. Actor (subject [person] of unit)
  1. Patient himself
  2. Patient's family
  3. Patient's spouse, date, fiancee
  etc.
- B. Time (of occurrence of described event)
  1. Present
  2. Past
  3. Future
  4. Subjunctive
- C. Interpersonal Acts (overt motor behavior involving two or more people, e.g., telling, hitting, leaving someone)

OR

- C'. Interpersonal States (nonovert states, thoughts, attitudes, etc., involving two or more people; e.g., thinking of, being angry at, or afraid of someone)
  1. Interpersonal System Categories to 16
  17. Neutral, unscorable

OR

- D. Noninterpersonal Acts (overt motor behavior involving only one person, e.g., eating, bathing, walking, etc.)

OR

- D'. Noninterpersonal States (nonovert states, thoughts, attitudes, etc., involving only one person, e.g., feeling ill, being sleepy, poor, happy, etc.)
  1. Positive (welcome, pleasant, etc. to patient)
  2. Neutral (neutral or indeterminant in significance to patient)
  3. Negative (disliked, unpleasant, etc. to patient)
- E. Object (person acted on or with, in interpersonal units)
  (As for A above)
- F. Topical Area (of general life experience)
  1. Financial
  2. Marital—Sexual
  3. Religious—Philosophical
  4. Educational
  etc.

II. Direct Interaction (units dealing directly with current interview interaction)
- A. Agrees, expresses compliance with interviewer
- B. Disagrees, expresses noncompliance with interviewer
- C. Asks for information, repetition, clarification, etc., from interviewer

etc.

III. Unscorable

described by Chapple[2] as involving both a physical high energy aspect *and* an out-going, interaction seeking aspect, it was further expected that one or more of them would be related to the degree of emphasis which the patient's content placed on interacting with others, Self C + C' in Table 2. Since the degree to which a patient takes the initiative in speaking, when the interviewer deliberately remains silent, has been proposed by Chapple as a measure of both the "drive" aspect of behavior and of the independence and scope of a person's interpersonal relationships, it was tentatively expected that Patient's (Pt.'s) Initiative would covary with content measures of breadth of interests (number of different persons and topics mentioned) and with description of the self in dominant interpersonal roles. It was such expectations as these, then, which guided the choice of dimensions of content to be included with the overly narrow Leary System in satisfying our goal of a comprehensive and multilevel content system.

The specific content scoring procedures include the stipulations that the categories are to be applied by the judges with a minimum degree of inference, and from the point of view of the patient. That is, all coding under this system is performed according to the relationship to or impact on the patient of the events as he describes them, without consideration of possible interpretations of psychological defenses or mechanisms, etc.

The content unit utilized for dividing verbalizations into countable and codable segments was basically defined as the minimal verbal statement which consensus of raters indicates to be understood as expressing an independent communication or thought. Although it was developed and tested independently, this unit is very similar to that of Auld and White (1956) and of Murray (1956).

Both the content unitizing and catgorizing processes have been shown to have adequate reliability when applied independently by trained judges according to detailed definitions and rules (Phillips, 1957).

Raw frequency scores (number of content units coded for each category) were converted into percentages so that intersubject comparisons would be unaffected by differences in absolute numbers of unitized items. The percentage scores were transformed by the arc-sine transformation (Snedecor, 1946) for purposes of statistical analysis. The major content scores were then correlated (Pearson *r*) with 12 temporal Interaction Chronograph variables. Means and other statistics, when obtained, were converted back to percentages.

RESULTS

Table 3 presents the major findings. In order to conserve space, only those relationships which reached statistical significance, and those approaching significance (given in pa-

[2] E. D. Chapple. *Manual for the Interaction Chronograph*, personal communication, undated.

TABLE 3

CORRELATIONS BETWEEN CONTENT CATEGORIES AND INTERACTION CHRONOGRAPH VARIABLES

Content	Interaction Chronograph						
	Pt.'s Units	Pt.'s Tempo	Pt.'s Silence	Pt.'s Adj.	Pt.'s Init.	Pt.'s Quick.	Pt.'s Domin.
Description (I)			−.42	(−.34)		−.40	
Self Subject (A)				.48			.36
Self Acts (D + C)	−.38			−.40			
Self Interpersonal (C + C′)	(−.34)						
Self Dominant-Hostile Quad.		.38					
Self Submissive-Hostile Quad.				(.33)			
Self Dominant-Positive Quad.							.37
Self Submissive				.41			
Self Positive							(.33)
Self Noninterpersonal (D + D′)	.37			(.35)			
Self Noninterpersonal Negative (D + D′ − 3)				−.36			
Other Interpersonal (C + C′)				−.38			
Other Noninterpersonal (D + D′)			(−.33)	−.46			.36
All Interpersonal (C + C′)	−.40						
Direct (II)	(.34)	(−.31)	.39	(.32)		.38	
Number Topics (F)	(−.31)		(−.31)		.46	−.36	
Number Persons (A + E)				−.52	(.28)	(−.34)	

Note.—$r = .36$ for $p = .05$. $r = .46$ for $p = .01$. Parenthetical values approach significance at .05 level.

rentheses), are included. Table 3 also does not include content categories which are essentially reciprocals of those presented, nor some Interaction Chronograph variables which have been shown to be highly correlated with those included and whose correlations with content categories essentially duplicated these results.[3]

Pt.'s Units. Because of the relatively fixed length of the interview and of the interviewer's utterances, Pt.'s Units is highly correlated in a negative direction with measures (Pt.'s Action, Pt.'s Tempo, and Pt.'s Activity) of the duration of the patient's utterances. In the present sample it correlates −.70 with Pt.'s Action and may be assumed to be one of the more stable and representative measures of the general level of the patient's verbal output because, as a *frequency* measure, it is less affected by a few extremely short or long utterances than are the duration measures.[4] The results shown in the first column of Table 3 reveal that the patient who has fewer Pt.'s Units, that is, who speaks relatively infrequently (but with longer durations per utterance), is shown by these correlations with content measures to describe himself as relatively: more active in his daily living, more oriented toward interactions with other people, less concerned with his own solitary experiences, interested in a wider variety of events in daily living, and less prone to evade description of himself in general.

Pt.'s Tempo. This is highly related to Pt.'s Units ($r = -.84$ in this sample) since the less frequently a patient speaks in the standardized interview, the longer *of necessity* will be his Tempo or "cycle" of speaking. Hence the relationships shown for Pt.'s Tempo should be considered in close conjunction with those for Pt.'s Units. As indicated in

[3] The complete matrices for content vs. Interaction Chronograph correlations have been deposited with ADI. Order Document No. 6641 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

[4] The fact that these highly related measures of verbal activity level do, however, at times differ somewhat in their relationships with external behavior points out the necessity for retaining all of the Interaction Chronograph variables despite some high intercorrelations, until the areas of their overlap and independence have been more completely defined and explored.

Table 3, the longer the average duration of a patient's Tempo, the more he describes himself as dominant-hostile in his dealings with and attitudes toward other persons.

Pt.'s Silence. This is negatively related to the percentage of content units which are *descriptive* in nature and is positively related to the relative frequency of direct interaction with the interviewer. One might hypothesize that both silence behavior and remarks made to the interviewer represent "resistance" or avoidance tactics, since the questions asked of the patients are nondirective ones calling for description of their general life patterns and do not deal with events within the interview. This hypothesis is also supported by the tendency noted above for the number of (short) utterances to be inversely related to such direct discussion with the interviewer, as well as by the trend for longer silences to be accompanied by the introduction of fewer different topics.

Pt.'s Adjustment. Since all patients hesitated before responding for longer durations than they interrupted, in analyzing the data, the minus sign was omitted, so that high scores for Pt.'s Adjustment indicate greater relative latency of response. The results in Table 3 show that patients with high Pt.'s Adjustment scores can be considered to be almost opposite to those with fewer (but longer) Pt.'s Units. Thus, "maladjustment" is *negatively* related to the relative frequency of description of the self as active and to the number of different other persons mentioned, while it is positively related to the relative amount of talk about the self. Further, the "slow responders" relatively less frequently describe the acts or the attitudes, both interpersonal and noninterpersonal, of *other* people; when they do describe their own interpersonal attitudes or dealings with other people, they relatively more frequently characterize themselves as taking a submissive role, probably one which is also hostile. They also mention fewer negative noninterpersonal things about themselves while tending to focus more upon the noninterpersonal aspects of their lives in general. Similar to Pt.'s Silence, this latency-of-response measure tends to be negatively associated with amount of descriptive content and positively related to relative frequency of direct interaction with the interviewer, supporting the hypothesis that this distinction between description and interaction with the interviewer represents avoidance in the face of difficulty in communicating.

The similar nature of the content category correlates of Pt.'s Units and Pt.'s Adjustment suggests what might be termed an "outward" or "other-directed" orientation in those patients who speak a fewer number of times and also in those who speak with a shorter latency of response. Patients who speak more times and those who wait longer before speaking have content which focuses more upon themselves, and thus might be termed "inward" or "self-directed." This interpretation is supported by our recent finding that schizophrenic patients, who might be thought of as being at the extreme pole on a continuum of inward-directed vs. outward-directed orientation, speak a significantly greater number of times (but in shorter average utterances), and with much longer average latencies before responding, compared to normals (Matarazzo & Saslow, 1961).

Pt.'s Initiative. The more the patient shows temporal interactional initiative (speaks again following his own last utterance), the more he also shows a form of "initiative" in raising new topics (and the broader one might therefore infer his interests or concerns to be). A similar but nonsignificant trend is shown for the total number of different persons mentioned.

Pt.'s Quickness. These results are supported by the finding that another temporal variable has similar correlations with the number of topics and number of persons mentioned. Unlike Pt.'s Initiative, however, Pt.'s Quickness is similar to Pt.'s Silence in being related significantly to the relative amount of descriptive content vs. interaction with the interviewer. Thus Quickness seems to have two components: one, the *readiness* to communicate in patients who take the initiative *rapidly*, a covariation which is similar to that found for Silence; and secondly, a component which, like the relationship found for Initiative, is related to the broader range of concerns (more people and more topics) of these same patients (with the shorter Quickness durations).

Pt.'s Dominance. This is directly related to the relative frequency with which the patient talks about people other than himself as well as to the relative concern he shows for the noninterpersonal feelings and behavior of others. Particularly striking is the finding that the more ***dominant*** the patient is in his temporal interview ***behavior***, the more he describes himself in ***content*** as dominant-positive (e.g., teaching, helping, advising, protecting, etc.) in his attitudes and dealings with other people. Thus it seems that for the patient whose interpersonal style is described (in content) as more dominant, the interruption behavior on the part of the interviewer in Period 4 is animating and challenging, rather than defeating. The findings for Pt.'s Dominance are very similar to those for Pt.'s Units, the content correlates of which also emphasized relatively more discussion of others. However, patients with higher verbal output, as defined by Pt.'s Tempo, more frequently described themselves as dominant-***hostile*** rather than dominant-***positive*** in their own interpersonal roles.[5]

Discussion and Summary

The results seem both encouraging in an exploratory study relating two quite disparate phenomena of interview behavior, and suggestive in the new meaningfulness which they add to the Interaction Chronograph variables by indicating relationships which seem inherently plausible and internally consistent. They provide a foundation for an approach to personality which combines content and temporal variables, and suggest personality dimensions which, underlying as they do at least two quite different spheres of behavior, may be particularly pervasive and consistent.

Viewed as a whole, the data presented in Table 3 suggest that patients who speak less often, who are faster to respond, and more dominant in the interview (that is, who have fewer and hence longer Units, shorter Pt.'s Adjustments, and more Dominance in the interruption stress period) have interview content which is relatively more oriented towards other people and towards interpersonal interaction, with social roles which are relatively more frequently described by them as dominant, either in a paternalistic or a hostile fashion. On the other hand, the correlations imply that the more a patient loses or submits to interruptions, is hesitant in speaking, and is less active verbally, the more his content emphasizes his own noninterpersonal concerns rather than interaction with others, and the more submissively hostile is his self-described role with other people.

In conjunction with the results of other validity-oriented studies of interview behavior correlates (Hare, Waxler, Saslow, & Matarazzo, 1960; Matarazzo, Matarazzo, Saslow, & Phillips, 1958; Matarazzo & Saslow, 1961), the present findings are a beginning at describing the characteristics which suggest a significant and cohesive description of the patient and how he interacts with others, viewed both from the subjective (content-inferred) and objective (temporally measured behavior) levels of observation. The major implication of these results bears upon the degree of generalizability of the Interaction Chronograph constructs and hence has to do with their concurrent (and, more remotely, construct) validity.

The correlations shown in Table 3 are small in the sense of accounting for relatively little of the variance, although respectable for complex personality variables. Further, the complete correlation matrix contained 336 r's (28 content categories vs. 12 Interaction Chronograph variables), of which 28 were significant at the .05 level or better. Since 17 would have been expected to be significant at this level by chance alone, a replication study is being undertaken to determine whether the present findings can be cross-validated with another sample of subjects. A number of hypotheses are suggested by the results of this validity-oriented study which can be pursued in future investigations if the present findings are borne out in the replication study.

[5] There is no relationship in this sample between Pt.'s Units and Dominance ($r = -.27$, not significant), while Dominance and Pt.'s Adjustment are negatively correlated ($r = -.45$). Pt.'s Adjustment is positively correlated with Pt.'s Units, however ($r = .45$).

References

Auld, F., Jr., & White, Alice M. Rules for dividing interviews into sentences. *J. Psychol.*, 1956, **42**, 273–281.

BALES, R. F. *Interaction process analysis.* Cambridge: Addison-Wesley, 1950.

CHAPPLE, E. D. The Interaction Chronograph: Its evolution and present application. *Personnel,* 1949, **25**, 295–307.

CHAPPLE, E. D., & ARENSBERG, C. M. Measuring human relations: An introduction to the study of the interaction of individuals. *Genet. psychol. Monogr.,* 1940, **22**, 3–147.

FREEDMAN, M. B., LEARY, T. F., OSSORIO, A. G., & COFFEY, H. S. The interpersonal dimension of personality. *J. Pers.,* 1951, **20**, 143–161.

GOLDMAN-EISLER, FRIEDA. Individual differences between interviewers and their effect on interviewees' conversational behavior. *J. ment. Sci.,* 1952, **98**, 660–670.

HARE, A. P., WAXLER, NANCY, SASLOW, G., & MATARAZZO, J. D. Simultaneous recordings of Bales and Chapple interaction measures during initial psychiatric interviews. *J. consult. Psychol.,* 1960, **24**, 193.

LENNARD, H. L. Concepts of interaction. Unpublished doctoral dissertation, Columbia University, 1955.

LUNDY, B. W. An investigation of the process of psychotherapy. Unpublished master's thesis, University of Chicago, 1950.

LUNDY, B. W. Temporal factors of interaction in psychotherapy. Unpublished doctoral dissertation, University of Chicago, 1955.

MATARAZZO, J. D., & SASLOW, G. Differences in interview interaction behavior among normal and deviant groups. In I. A. Berg & B. M. Bass (Eds.), *Conformity and deviation.* New York: Harper, 1961, in press.

MATARAZZO, J. D., SASLOW, G., & HARE, A. P. Factor analysis of interview interaction behavior. *J. consult. Psychol.,* 1958, **22**, 419–429.

MATARAZZO, J. D., SASLOW, G., & MATARAZZO, RUTH G. The Interaction Chronograph as an instrument for objective measurement of interaction patterns during interviews. *J. Psychol.,* 1956, **41**, 347–367.

MATARAZZO, RUTH G., MATARAZZO, J. D., SASLOW, G., & PHILLIPS, JEANNE S. Psychological test and organismic correlates of interview interaction patterns. *J. abnorm. soc. Psychol.,* 1958, **56**, 329–338.

MURRAY, E. J. A content-analysis method for studying psychotherapy. *Psychol. Monogr.,* 1956, **70**(13, Whole No. 420).

PAGE, H. A. An assessment of the predictive value of certain language measures in psychotherapeutic counseling. In W. V. Snyder (Ed.), *Group report of a program of research in psychotherapy.* University Park, Pa.: Pennsylvania State University, 1953.

PHILLIPS, JEANNE S. The relationship between two measures of interview behavior comparing verbal content and verbal temporal patterns of interaction. Unpublished doctoral dissertation, Washington University, 1957.

SASLOW, G., & MATARAZZO, J. D. A technique for studying changes in interview behavior. In E. A. Rubinstein & M. B. Parloff (Eds.), *Research in psychotherapy.* Washington, D. C.: American Psychological Association, 1959.

SNEDECOR, G. W. *Statistical methods.* (4th ed.) Ames: Iowa State College, 1946.

(Received May 27, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 267-269

THE CONCEPT OF NORMALITY:
A REPLY TO FREIDES

MAURICE KORMAN
University of Texas Southwestern Medical School

In an interesting paper published in this journal, Freides (1960) presses for the elimination of the concept of normality on the grounds that there is little agreement concerning its definition, that it ascribes absolute, yet culture bound, patterns of behavior and that it disregards the flexible interactions between personality and circumstances. Although Freides concerns himself nearly exclusively with the idealist-adjustment view of normality, other approaches such as the statistical-average conception are likewise dismissed with the conclusion that "for purposes of scientific theory and also for practical clinical purposes" our emphasis must shift away from considerations of normality (or pathogenicity) and toward a greater concern with "the potentialities of every person under the proper conditions." Such a position, it will be argued here, springs from too narrow a conceptualization of both scientific theory and clinical purposes. It has, moreover, some unfortunate implications for future research developments and basic orientations in clinical psychology.

Although normality is (today) an admittedly low powered concept, judgments of normality-pathogenicity hover in the background of many diagnostic and therapeutic decisions. It seems that, however ambiguous and ill-defined a conception of normality our professional behavior reflects, this term nonetheless behaves the way a good construct should (Beck, 1953): as a *summarizer* of a host of personality characteristics and behavioral tendencies, as a gross *predictor* of a person's future behavior and as an *object of search*. This last role will be made clearer as this discussion develops.

If our clinical practice reflects (as I believe it does) the use of *some* concept of normality, it becomes appropriate to inquire into what kind of meaning can be ascribed to it. The controversy over the idealist-adjustive vs. the statistical-average interpretations of normality has obscured some more basic considerations. Why should the question, "What is normality?" have a different logical status than, say, the question, "What is schizophrenia?" That a diagnosis of schizophrenia, for instance, implies a hypothesization of an inner structure or state which is but inadequately described by an operational definition, and that such taxonomic conceptualization is a scientifically valuable activity has been convincingly argued by Meehl (1959). Although our present day conception of schizophrenia is riddled through with indeterminacies, we would be hard put to do without it—clinically or scientifically. The concept of normality suffers from the same type of inadequacies, only more so, since it belongs to a far less imposing nomological network than does schizophrenia.

In the case of both schizophrenia and normality we are asking legitimate questions to which, unfortunately, no sufficiently complete or specific answer can be given *at this time*. But scientific theory as it is understood today (Hempel, 1952) accepts vagueness or open endedness in the real definitions (i.e., definitions involving a statement of the essential nature or characteristics) of concepts. We should understand the question, "What is normality?" not as a demand for an unequivocal definition (as Freides implicitly does) but as a request for an empirically based specification of the *indicators* (prevalence of control factors? level of corticoseptal integration? degree of reasonableness? appropriateness of autonomic arousal? etc.) which may be found

to reflect with varying probability the existence of this hypothetical state. During the early stages of a science, the specification of the meaning of concepts is, as Kaplan (1946) notes, "a provisional one, both as to the indicators included and the weights associated with them."

The definition of normality, the hunt for the indicators of its essential characteristics, is thus construed as a processive affair marked by high responsivity to new data. An illustration incorporating some suggestive data might be of interest. To begin with, let us consider a first temporary indicator of normality such as nonexposure to psychiatric diagnosis and treatment. As an operational definition of normality, this criterion would obviously be grossly inadequate as would many others. Our gambit rests on the hope that it may eventually enable us to lift ourselves by our bootstraps (Cronbach & Meehl, 1955) by calling attention to other "purer" criteria which correlate (imperfectly) with it and which may in turn possess greater validity than our original indicator does. Many of our more successful constructs in psychology (e.g., Binet IQ) have evolved in this way. Using such a first indicator we would orient ourselves to discovering what other indicators (and characteristics by implication) compose a matrix of significant relatedness. It is clear at this point that this approach, in contrast to the idealist or average views of normality, would lead us to an empirically based conceptualization of normality having the essential character of a theoretical construct. We might pause to ask, how much reliable data do we happen to have concerning these "nonexposed" (or otherwise discriminated) normals? Surprisingly little. From reading the psychological literature a Martian might be led to believe that the proportion of schizophrenics in the United States is 80% instead of .8% or that the nonexposed represent only some 10% of the population when they should number closer to 90%.

A perusal of the literature dealing with such loosely defined normals leaves one major impression: there is an unexpectedly high distribution of "pathogenic" traits, histories, behaviors, in this population while by contrast many of the allegedly normal (ideal-adjusted) characteristics we have been taught to expect are by no means typical of it (Lapousse & Monk, 1958; Renaud & Estess, 1955; Schofield & Balian, 1959). For our purposes this would suggest that quite a few of the admirable, clearly nonpathogenic traits and behaviors with which we have traditionally invested normality relate poorly to this concept as defined (preliminarily) by the nonexposure and other similar indicators. Parenthetically, this represents an attack on the same idealist-adjusted definition of normality that Freides finds unsatisfactory—*without, however, eliminating normality as an object of scientific search.* On the contrary, research on normality should be quickened and stimulated by such findings, presuming they stand up under more systematic investigation. Are there, for instance, suppressor-control variables which override the effect of such pathogenic events (Schofield & Balian, 1959), and are *they* the indicators of normality? Or will the answer involve a favorable interaction between certain constitutional factors and patterns of personal history? We certainly do not know the answer at this time. There can be little doubt, however, that the concept of normality constitutes an object of search that is not only scientifically appropriate but also of great moment to psychology. Such research emphasis may well be a factor in helping to precipitate a much needed shift away from the basic orientation to pathology that characterizes many of our efforts today. Clinical psychology needs to reacquaint itself with its most unique and natural subject matter—normal man. It is unfortunate, as Sanford (1958) has pointed out, that "we . . . talk much more freely about symptoms of illness than about symptoms of health or symptoms of resiliency or of strength or of spontaneity or of creativity" (pp. 82–83).

Little has been said in this rejoinder concerning the aspect of the problem of normality which bears Freides' strongest criticisms, the evaluative component of normality. Our approach simply sidesteps this argument altogether. Normality as here conceived involves absolute values or cultural biases *no more and no less* than does, e.g., our concept of schizophrenia. Certainly normality is culture bound, but only to the extent that all concepts are "bound" to the phenomena they describe or

explain. Likewise, what kind of value judgment would be involved in the empirical determination that, e.g., a specific suppressor-control variable constitutes an essential characteristic of normality? As far as we can see the only values involved in this enterprise are the same that motivate all scientific research and as such they are rightly accepted as neutral with reference to the content of science.

Although the bulk of this paper has been devoted to a defense of a conceptualization of normality, we might briefly consider the alternative that Freides champions, i.e., that we concern ourselves instead with the specific abilities and limitations that individuals demonstrate under specific circumstances. This point of view has a long history in psychology as the specificity theory of personality, and as such has been much discussed over the years. Some general comments, however, seem appropriate here. First, in addition to seeking the highest possible accuracy in the explanation and prediction of behavior, science also attempts to be maximally comprehensive and parsimonious. Freides' approach represents an admission of failure with respect to these desiderata. Secondly, the two approaches are by no means incompatible. Classifying a person as "normal" does not by any stretch of the imagination preclude an appreciation of his limitations under certain circumstances. It can easily be conceived that broadly specified situational variables may play a significant role in behavior prediction equations. To what extent this is true we do not know. It is further possible, once this typological approach has begun to yield some fruits, that we will be able to shift to systematic studies of the *processes* which favor normal psychological growth. It is clear, however, that psychology would be poorly served if we gave up our search for the meaning of normality—a potentially powerful explanatory and predictive construct.

REFERENCES

Beck, L. W. Construction and inferred entities. In H. Feigl & May Brodbeck (Eds.), *Readings in the philosophy of science*. New York: Appleton, 1953.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

Freides, D. Toward the elimination of the concept of normality. *J. consult. Psychol.*, 1960, **24**, 128–133.

Hempel, C. G. Fundamentals of concept formation in empirical science. *International encyclopedia of unified science*, Vol. II, No. 7. Chicago: Univer. Chicago Press, 1952.

Kaplan, A. Definition and specification of meaning. *J. Phil.*, 1946, **43**, 281–288.

Lapousse, Rema, & Monk, Mary A. An epidemiologic study of the behavior characteristics in children. *Amer. J. publ. Hlth.*, 1958, **48**, 1134–1144.

Meehl, P. E. Some ruminations on the validation of clinical procedures. *Canad. J. Psychol.*, 1959, **13**, 102–128.

Renaud, H., & Estess, F. Life history interviews with 100 normal American males: "Pathogenicity" of childhood. *Amer. Psychologist*, 1955, **10**, 371.

Sanford, F. H. Psychology and the mental health movement. *Amer. Psychologist*, 1958, **13**, 80–85.

Schofield, W., & Balian, Lucy. A comparative study of the personal histories of schizophrenic and nonpsychiatric patients. *J. abnorm. soc. Psychol.*, 1959, **59**, 216–225.

(Received June 6, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 270–271

THE EFFECTS OF TWO VERBAL TECHNIQUES ON THE EXPRESSION OF FEELINGS[1]

GUSTAV LEVINE[2]
Teachers College, Columbia University

A problem frequently encountered in psychotherapy is the client's use of impersonal, nonemotional statements. The therapist's requests for statements referring to the client's feelings are frequently met with descriptions of situations or impersonal observations, even though the client is highly motivated to cooperate. Therapists therefore frequently attempt to behave in a manner that facilitates their client's emotional expression.

Rogers, in his early writings, stated his observation that reflection of feelings results in the immediate expression of further feelings by the client (Rogers, 1942, p. 158). Studies of recorded therapy sessions which utilized categories of clients' expressions of feelings and therapists' reflections of feelings, have not examined the specific sequence of reflection and expression of feelings (Bergman, 1951; Seeman, 1949; Snyder, 1945). Verplanck (1955), in an experiment involving a social rather than a therapeutic situation, found that paraphrasing, a technique which is descriptively similar to reflection, increased the class of statements paraphrased. Although feelings were not paraphrased, but rather statements of opinion, the results are encouraging to a hypothesis that reflection of feelings increases expression of feelings. In the Verplanck study, paraphrasing of the class of responses to be increased was the only response given by the interviewers. This restricted attention to one class of response may have been the factor which reinforced this response (rather than the specific paraphrasing technique), implying that any technique involving restricted attention to one class of response would reinforce that class. On the other hand, the specific technique (paraphrasing) may be reinforcing in the same way that approval can be reinforcing (Murray, 1956). The contention that restricted attention is the significant factor is supported by the work of Salzinger and Pisoni (1958, 1960), who found that simple statements such as, "I see," "Yeah," "Uhha," "Mmm-hmm," etc., could act as positive reinforcers of expression of feelings.

The present study involves a comparison of the specific technique of reflection of feelings with a simple undifferentiated vocalization, "Mm-hm," both applied as the only response given, in separate interviews, and occurring only after expressions of feeling. In such a pair of interviews both interviewer techniques would constitute restricted attention to a class of response, and any special advantage of the more complex response could express itself in more effective reinforcement.

It was hypothesized that reflection of feelings results in greater expression of feelings than does an undifferentiated vocalization ("Mm-hm").

The subjects were 30 male undergraduate students, living at college, who volunteered and were paid for their participation.

The experiment was explained as a study of feelings about school life. The experimenter stated that he was "interested in what you are experiencing" as a student. The subjects each received two similarly structured interviews, which differed only in the way in which the experimenter responded to expression of feel-

[1] This paper is taken from portions of a thesis submitted to the Department of Clinical Psychology, Teachers College, Columbia University, in partial fulfillment of the requirements for the PhD degree. The author wishes to express his appreciation to his Chairman, Joel Davitz, for his aid and encouragement.

[2] Now at the Creedmoor Institute for Psychobiologic Studies.

ings by the subjects (reflection in one, "Mm-hm" in the other).

The interviews were scheduled one week apart and tape recorded. One half of the subjects received one condition first, the other half receiving the other condition first.

The interviews were rerecorded with the interviewer's responses omitted, and with the tape divided into one-minute segments of talk by the subjects. The presence or absence of expression of feelings in a single minute was determined by six psychologists. These judges used a detailed set of instructions given to them by the experimenter as a frame of reference for their judgments.[3]

The basic data for the testing of the hypothesis was the number of minutes containing expressions of feeling.

The one-minute segments were rated for presence or absence of feeling by both the experimenter and a judge. The agreement was high, a phi coefficient of .92 having been obtained through a transformation of the computed chi square value of 428.9.

The average number of minutes of feeling in each of the two interview conditions was computed. In the reflection interviews there was an average of 10.07 minutes containing expression of feelings, and in the "Mm-hm" interviews there was an average of 9.83 minutes containing expression of feelings. A t test of the difference yielded a nonsignificant t of .23.

The results indicate that under the conditions of the present experiment there is no difference in effectiveness between reflection of feelings and the undifferentiated vocalization "Mm-hm," as techniques for increasing expression of feelings. The additional factor of clarification of feelings through re-expression in different words does not seem to increase the frequency of expression of feelings beyond that obtained with any technique which responds only to feelings.

The experiments of Verplanck (1955) and Greenspoon (1955), indicate that the two techniques can each be effective compared to no technique (operant level), but there was no control in the present experiment with which to make such a comparison.

[3] The instructions are in the appendix to the thesis. The thesis can be obtained on microfilm from University Microfilms: 313 North First Street; Ann Arbor, Michigan. Order L.C. Card No. Mic. 58-2237, remitting $2.00.

REFERENCES

BERGMAN, D. Counseling method and client responses. *J. consult. Psychol.*, 1951, **15**, 216–224.

GREENSPOON, J. The reinforcing effect of two spoken sounds on the frequency of two responses. *Amer. J. Psychol.*, 1955, **68**, 409–416.

MURRAY, E. J. A content-analysis method for studying psychotherapy. *Psychol. Monogr.*, 1956, **70**(13, Whole No. 420).

ROGERS, C. R. *Counseling and psychotherapy*. Boston: Houghton Mifflin, 1942.

SALZINGER, K., & PISONI, STEPHANIE. Reinforcement of affect responses of schizophrenics during the clinical interview. *J. abnorm. soc. Psychol.*, 1958, **57**, 84–90.

SALZINGER, K., & PISONI, STEPHANIE. Reinforcement of verbal affect responses of normal subjects during the interview. *J. abnorm. soc. Psychol.*, 1960, **60**, 127–130.

SEEMAN, J. A study of the process of nondirective psychotherapy. *J. consult. Psychol.*, 1949, **13**, 157–168.

SNYDER, W. U. An investigation of the nature of nondirective psychotherapy. *J. gen. Psychol.*, 1945, **33**, 192–223.

VERPLANCK, W. S. The control of the content of conversation: Reinforcement of statements of opinion. *J. abnorm. soc. Psychol.*, 1955, **51**, 668–676.

(Received April 29, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 272

BRIEF REPORTS

ANXIETY IN VERBAL BEHAVIOR:

AN INTERCORRELATIONAL STUDY [1]

MERTON S. KRAUSE
University of Michigan

Several objective measures of anxiety in verbal behavior have been proposed in the last 3 decades. Their claims to validity are relatively weak, generally resting upon their apparent reasonableness. If they appear to reflect the same "inner state" and so yield highly correlated measurements on the same behavior samples, their claims would be stronger. They might be said to show concurrent validity.

Ten-minute recorded samples from the therapy sessions of 15 hospitalized male mental patients were studied. Eight purported measures of anxiety were applied to each patient's verbal response in the 15 protocols. The measures were (*a*) number of words spoken, (*b*) number of words/number of inspirations, (*c*) number of verbs/number of adjectives, (*d*) latency of the response, (*e*) number of references to the interviewer, (*f*) and (*g*) number of speech disruptions as measured by Mahl (1956) "non-ah" ratio and Dibner (1956) Cue Count I and (*h*) rate of speech.

An intercorrelation matrix was computed for each of the 15 protocols to study the amount of individual differences, and then the matrix of median intercorrelations was derived. The average level of correlation in this later matrix was .06, and so no general convergence of the several measures is evident. If these measures do possess some degree of concurrent validity it was not uniform enough over persons to appear in our sample. The cluster pattern in the matrix of median intercorrelations does suggest some convergences in our set of measures. As might be expected by inspection of the measures themselves, the two speech disruption measures were highly correlated ($r_{67} = .91$), while verbs/adjectives and number of words were moderately loaded on the same factor (about .42 and .24, respectively). This pattern of Measures *f*, *g*, *c*, and *a* held up for about half of the protocols. In at least two protocols, however, this pattern clearly disintegrated.

The average results have a very large sampling error over persons. The average interquartile range for the median values was .45. Thus, there are marked individual differences in the level of intercorrelation among the various verbal anxiety measures. This implies that different measures may be valid for different persons and that what measurement values indicate anxiety or nonanxiety may also be idiosyncratic. These results suggest that verbal measures are not going to be any less troublesome to validate than are physiological measures of anxiety.

[1] This study was supported under Grant M-516 C-7 from the National Institute of Mental Health, E. S. Bordin, Principal Investigator.

An extended report of this study may be obtained without charge from Merton S. Krause (2343 Auburn Avenue; Cincinnati 19, Ohio) or for a fee from the American Documentation Institute. Order Document No. 6642, from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

REFERENCES

DIBNER, A. S. Cue counting: A measure of anxiety in interviews. *J. consult. Psychol.*, 1956, **20**, 475–478

MAHL, G. F. Disturbances and silences in the patients' speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, **53**, 1–15.

(Received July 1, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 273

VERBAL AND PERCEPTUAL COMPONENTS IN WISC PERFORMANCE AND THEIR RELATION TO SOCIAL CLASS [1]

JOHN B. MARKS
Veterans Administration Hospital, American Lake, Washington
AND JAMES E. KLAHN
Tacoma Public Schools

Most investigators have found measured intelligence positively related to social class and have found this relation closer with verbal rather than nonverbal materials. With the WISC, however, Estes (1953) found higher status superiority only at a 7-year-old level and not at the 10-year-old level. Moreover, she found no consistent pattern differences of subtests between her upper social group and her lower group.

The present study relates social class to two WISC measures of verbal-nonverbal difference: (*a*) Verbal IQ-Performance IQ, and (*b*) the difference between weighted scores of subtests highly loaded on Cohen's (1959) verbal factors and weighted scores of subtests highly loaded on the perceptual factor. This latter was the mean of Information, Comprehension, Similarities, and Vocabulary minus the mean of Picture Completion, Block Design, and Object Assembly.

Subjects were 211 primary school children divided by age and sex into four groups. Both younger groups were within 6 months of their eighth birthday at testing while the older groups were within 6 months of their eleventh birthday. All had been tested because of some school difficulty but children with IQs below 70 were eliminated in sample selection. Mean IQs of the sample were close to population means.

From information about the father's occupation each subject was assigned to an occupational class group ranging from 1, casual laborer, to 9, business leader. The r between independent ratings of the two authors was .94 and means of the two ratings were used.

Occupational ratings correlated positively with IQs in both the younger and the older groups. Though the correlation in the younger group is higher it is not significantly so. On the other hand, the girls show a substantially higher correlation than do the boys. For verbal and full-scale IQs this difference is significant, .42 compared to .19 for verbal and .45 to .17 in the full-scale.

The two measures of verbal-nonverbal difference were tested for their relation to occupational level, both directly through correlations, and by contrasting the difference measures for unskilled and semiskilled labor children with those for white collar children. The difference between verbal and performance IQs was in the expected direction but not significant. The difference in the factor-derived measures had a significant r of .16 ($p < .05$, $N = 211$) with occupational level and the white collar children showed a greater verbal superiority ($t = 2.41$, $p < .01$, 113 df).

These results are consonant now with results using other instruments. On the WISC both younger and older groups show a correlation of IQ with occupational class, and this correlation is higher when verbal materials are used than it is for nonverbal. The closer relation between occupational level and IQ among girls than among boys may stem from the higher peer value which girls put upon middle class verbality.

[1] An extended report of this study may be obtained without charge from John B. Marks (Mental Health Research Institute; Fort Steilacoom, Washington) or for a fee from the American Documentation Institute. Order Document No. 6643, from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

REFERENCES

COHEN, J. The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *J. consult. Psychol.*, 1959, **23**, 285–299.

ESTES, BETSY W. The influence of socioeconomic status on the WISC: An exploratory study. *J. consult. Psychol.*, 1953, **17**, 58–62.

(Received July 26, 1960)

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 274

SENSORY DEPRIVATION AND ITS RELATION TO PROJECTION[1]

MALCOLM H. ROBERTSON AND ROBERT C. MARTIN
University of Florida

The study was designed to test the hypothesis that sensory deprivation lowers the threshold for projection. To test this hypothesis, a control group and an experimental group of 10 subjects each were used, half male, half female. The control subjects received no deprivation and were tested individually for projection using a modified autokinetic technique. The technique consisted of presenting the subject with a dim source of light approximately 1 mm. in diameter at a distance of 9 feet. The subject was told to watch the moving pinpoint of light and, when it went off, to report what it suggested, or looked like, or made him think of. There were 12 1-minute trials with a 2-minute rest period after the sixth trial.

The experimental subjects were exposed to sensory deprivation for 3 hours and then tested immediately with the autokinetic technique in the same manner as the control subjects. In the deprivation condition, each subject wearing opaque goggles, cotton mittens, and cardboard cuffs, was placed on a bed with his head inside a foam rubber lined box.

The two groups were compared in terms of the number of responses as well as the number of stimulus-bound responses, original responses, and popular responses. Stimulus-bound responses were those that referred solely to the movement or direction of movement of the light. An original response was one that was given only once, by only one person, and in addition was considered by the two investigators to be very novel or unusual. A popular response was one that was found in several records, usually occurring more than once in a record.

It was expected that the deprivation group would show more projection (greater productivity, larger number of original responses, and fewer popular and stimulus-bound responses) than the control group. Differences between the two groups were not statistically significant.

(Received August 8, 1960)

[1] An extended report of this study may be obtained without charge from Malcolm H. Robertson (Department of Psychology, University of Florida; Gainesville, Florida) or for a fee from the American Documentation Institute. Order Document No. 6644 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.25 for microfilm or $1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 275

PSYCHIATRIC OUTPATIENT PERSONALITY PATTERNS[1]

MARY HELEN TATOM

Spring Grove State Hospital, Baltimore

In a previous study (Tatom, 1958) five patients were selected on the basis of medical diagnoses as representatives of each of four commonly occurring nosological entities: obsessive-compulsive, hysteric, anxiety state, and outpatient schizophrenic. They were rated by their respective individual therapists on each of 67 personality variables, and the resulting scores intercorrelated. The 20 × 20 matrix of person-to-person correlations was factored by Thurstone's centroid method to yield four primary and two second-order factors, tentatively identified with clinical syndromes.

To test the stability of these patterns, 2 reference individuals for each of the four primary factors were inserted into a matrix with 16 new patients, selected to represent equally the original four clinical diagnostic entities. A factor analysis paralleling the first was carried out. The position of reference individuals with respect to primary factors in the two analyses is given in Table 1. Trait patterns of structurally corresponding factors in the two analyses suggested that syndromes were replicated in the second analysis, though overlap of specific items was not extensive (e.g., two internally consistent clusters of clinically schizoid traits had in common only *seclusiveness* and *poor social adjustment*). Primary factors were tentatively identified with the respective factors of the original analysis as follows: *outpatient paranoid schizophrenic, conversion hysteric, socially mature personality,* and *passive-dependent personality* vs. *antisocial personality.* The latter was less clearly defined by reference individuals than were the other three. Patients in both analyses were grouped largely on the basis of second order factors, rather than highly correlated primary factors. Factor and trait patterns tended to confirm two second-order factors: *schizothymia* vs. *cyclothymia,* and *uncontrolled emotionality* vs. *overcontrolled emotionality.*

Psychiatric and factorial classification of patients did not agree in either analysis; syndromes isolated factorially corresponded only roughly to the nosological entities represented by patients in both groups of patients analyzed.

[1] An extended report of this study may be obtained without charge from Mary Helen Tatom (641 Houston Ave., Apartment 202; Takoma Park, Maryland) or for a fee from the American Documentation Institute, Order Document No. 6645 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $2.00 for microfilm or $3.75 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

REFERENCE

TATOM, MARY H. A factorial isolation of psychiatric outpatient syndromes. *J. consult. Psychol.,* 1958, *22,* 73–81.

(Received August 25, 1960)

TABLE 1

PROFILES OF FACTORS IN REFERENCE INDIVIDUALS

Reference Individual	Factor							
	A	A_2	B	B_2	C	C_2	D	D_2
O-5 (A)	.60	.66	−.20	.00	.02	−.08	−.26	.13
O-3 (A)	.67	.87	−.03	.06	.00	−.05	−.05	−.07
A-1 (B)	−.20	−.27	.49	.71	.11	.28	−.01	.18
A-4 (B)	.00	−.02	.57	.64	.01	−.01	.04	−.14
H-3 (C)	.02	−.02	.01	.06	.69	.71	−.03	.00
A-2 (C)	−.08	−.39	.44	.01	.58	.40	.26	.51
H-2 (D)	−.26	−.64	−.11	.38	−.02	.39	−.49	.37
H-4 (D)	.13	−.39	.00	.52	−.04	.42	−.49	.23

Journal of Consulting Psychology
1961, Vol. 25, No. 3, 276

IDENTIFICATION IN TERMS OF PERSONAL CONSTRUCTS:

RECONCILING A PARADOX IN THEORY [1]

ROBERT E. JONES

Veterans Administration Hospital, Danville, Illinois

In this study, identification is defined as perceived similarity of self and others, experienced in terms of personal constructs. The author accepts the definition of phenomenological psychologists and the thinking of psychoanalysts, such as R. P. Knight, who emphasize identification as a relationship rather than a process, and as a perceived rather than as an actual similarity.

The psychology of personal constructs, the theoretical system developed by G. A. Kelly, is a perceptual approach to the prediction and explanation of human behavior. One's personal constructs are the vehicles, verbally expressed, by which one anticipates the behavior of others and guides his own behavior. Constructs are dimensions defined by terms which the perceiver accepts as opposites or contrasts. The way in which two persons are seen as alike yet different from a third would be a construct. A form of the Role Construct Repertory Test is employed to measure identification with "significant others."

Identification with male figures in the Repertory test was taken as the central measure in the present research. Repeat-test reliability was established in the high .80s. The subjects were 36 hospitalized males with mild or moderate psychiatric disorders, and a control group of 36 normal males, matched on age and education.

The central hypotheses were two: (*a*) neuropsychiatric (NP) patients more often than normal adult males will either overidentify or underidentify with personally significant male figures; and (*b*) the personal construct matrices of NP patients will be simpler than those of controls.

The hypothesis of underidentification is supported only at the 10% level, by one-tailed *t* test. The hypothesis of overidentification is significant at the 1% level: NPs are more likely to see others as extremely like the self than are normals. Also as predicted, the idiographic factors required to "explain" the interpersonal matrix are significantly fewer (at the .05 level) for NPs than for the controls. For NPs, but not for normals, the more simple the factor matrix the more fully it is explained by a value construct (.01 chi square significance).

The major contribution of the Personal Construct approach to identification theory lies in the reconciliation of previous theories. Both E. H. Erikson and O. H. Mowrer are partly right. Both over- and underidentification are common badges of maladjustment. Both are associated with a common defect—a factorially simple, value-laden system of constructs. Dimensions of perception permitting useful discrimination become inoperant. With the construct system compelling polarization into "good guys and bad," we see, with Sanford, how "desperation" promotes classical identification of the all-or-none variety. Identification with a hated object tends to be unconscious and accomplished by introjection. Identification with an idealized object tends to be conscious and achieved by assimilative projection. Sanford's "identification proper" is the unconscious type, always desperation motivated. Either type of identification, whether excessive or deficient, can be explained in terms of an oversimplified construct system, preempted by the value dimension.

The findings support Rinder and Campbell's contention that both over- and underidentification reflect the same neurotic dynamic: in their terms, "undue reaction-sensitivity"; in ours, undue perceptual restriction to the value dimension.

(Received September 22, 1960)

[1] An extended report of this study may be obtained without charge from Robert E. Jones (Psychology Service, Veterans Administration Hospital; Danville, Illinois) or for a fee from the American Documentation Institute. Order Document No. 6646 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance $1.75 for microfilm or $2.50 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

This paper draws on data collected when the author was engaged in doctoral research at Ohio State University.

PSYCHIATRY

JOURNAL FOR THE STUDY OF INTERPERSONAL PROCESSES

CONTENTS OF NO. 1, VOL. 24
FEBRUARY, 1961

RAYMOND FIRTH: **Suicide and Risk-Taking in Tikopia Society**

ROBERT E. NIXON: **An Approach to the Dynamics of Growth in Adolescence**

HAROLD I. LIEF and WILLIAM C. THOMPSON: **The Prediction of Behavior from Adolescence to Adulthood**

WALKER PERCY: **The Symbolic Structure of Interpersonal Processes**

CHARLES SAVAGE: **Countertransference in the Therapy of Schizophrenics**

WILLIAM N. DEANE: **The Reactions of a Nonpatient to a Stay on a Mental Hospital Ward**

EDWIN A. WEINSTEIN and ROBERT L. KAHN: **Patterns of Sexual Behavior Following Brain Injury**

BRIEF COMMUNICATIONS:

RAMON FERNANDEZ-MARINA: **The Puerto Rican Syndrome: Its Dynamics and Cultural Determinants**

BOOK REVIEWS

1610 New Hampshire Ave., N.W.,
Washington 9, D.C.

Subscription price $10.00, foreign $10.80

PSYCHOLOGY AND REHABILITATION

BEATRICE A. WRIGHT, Ph.D.

Editor

Based on the proceedings of an Institnte on the roles of psychology and psychologists in rehabilitation, sponsored by the American Psychological Association and supported by a grant from the Office of Vocational Rehabilitation, Department of Health, Education and Welfare.

The volume examines the principles and assumptions underlying the modern philosophy of rehabilitation, paying special attention to the roles and functions of psychologists as members of an interdisciplinary team. Problems of training are explored in detail and in the context of the changing nature of the field. Areas of needed research are highlighted.

Price $1.50

Order from:

American Psychological Association, Inc.

Dept. PAR

1333 Sixteenth Street, N.W.

Washington 6, D.C.

A NEW TEST!!

"The new in tests. . . . . ."

How would you like to use a projective personality test which can be scored quickly and objectively? Whose interpretation is based on careful factor-analytic studies? Which has norms for 11-year-old through adult levels (over 1900 selected *Ss*)? Uses an interesting and disarming cartoon-like format?

Cassel and Kahn

GROUP PERSONALITY PROJECTIVE TEST

Order Examiner's Set including general manual, scoring keys, 10 test booklets, and 100 answer sheets, enough material for a good-sized testing program (and only $13.50).

For further information, write to

PSYCHOLOGICAL TEST SPECIALISTS
Box 1441
Missoula, Montana

"The new in tests. . . . . ."

A NEW TEST!!

When the Mountain Did Go to Mahomet

For the parents of retarded or emotionally disturbed children, transportation expenses for enrollment and visits at a usually distant treatment center — added to the cost of residential treatment itself — could, on occasion, make it impossible for the parents to give their child the benefits of an individualized, twenty-four-hours-a-day program, under full professional guidance.

Cognizant of this factor, The Devereux Foundation has pioneered three branches, which, in effect, make it one of the most accessible residential treatment centers in the United States. At each branch outstanding therapeutic, educational, and vocational services are available.

Serving the East Coast, Devereux Schools are located at Devon, Pennsylvania (Mr. Charles J. Fowler, Director of Admissions); serving the West Coast, Devereux Schools at Santa Barbara, California (Mr. Keith A. Seaton, Registrar); and serving the Southwest, Devereux Schools at Victoria, Texas (Mr. John M. Barclay, Director of Development). These will welcome your inquiry.

THE DEVEREUX FOUNDATION

A nonprofit organization
Founded 1912
Devon, Pennsylvania
Santa Barbara, California
Victoria, Texas

SCHOOLS
COMMUNITIES
CAMPS
TRAINING
RESEARCH

HELENA T. DEVEREUX
Administrative Consultant

EDWARD L. FRENCH, Ph.D.
Director